Synthèse Probabilités et Statistiques

INFO B223

Berg Lucas - Benoit Gabriel

Année académique 2020-2021

Chapitre 1 : Statistique descriptive univariée

Définitions générales

Population : C'est l'ensemble des individus sur lesquels on désirer réaliser une ou plusieurs mesures d'un *caractère* (encore appelé *variable*).

Statistique : C'est une science dont l'objectif est d'interpréter les données récoltées au sujet d'une population ou d'un échantillon de cette population. Cela désigne également un ensemble données, on parle encore de série statistique associée à un caractère particulier.

Stocker une série statistique dans R : On va utiliser la classe *data.frame* et des *vecteurs*.
(Pour lire et écrire des fichiers xls, il faut la librairie *xlsReadWrite*)
Pour lire un fichier et le stocker dans une variable v, on peut utiliser les commandes :

- read.csv("/path/to/file", sep?=)
  Exemple : `v<-read.csv("/path/to/file", sep="")`
- read.table("/path/to/file", sep?=)
  Exemple : `v<-read.table("/path/to/file")`
- read.xls("/path/to/file")
  Exemple : `v<-read.xls("/path/to/file")`

Pour écrire dans un fichier, on peut utiliser les commandes :

- write.csv(données, "/path/to/file", sep?=)
- write.table(données, "/path/to/file", sep?=)
- write.xls(données, "/path/to/file")

Pour créer soit-même un data.frame et le stocker dans *frame*, on peut utiliser :

- des vecteurs
- la commande data.frame(vect, vect, ..., row.name?=vect)

Exemple :

```
age<-c(40, 50, 10)
taille<-c(150, 100, 180)
nom<-c("Billy", "Djordan", "Pepito")
frame<-data.frame(age, taille, row.name=nom)
```

Créer un vecteur et le stocker dans R :
Pour stocker un vecteur dans une variable v, on peut utiliser :

- l'opérateur " :"
  Exemple : `v<-1 :5` donnera `1 2 3 4 5`
- la commande seq(from=, to=, by=)
  Exemple : `v<-seq(from=1, to=5, by=2)` donnera `1 3 5`
- la commande rep(x=, times=)
  Exemple : `v<-rep(x=1, times=5)` donnera `1 1 1 1 1`
- la commande c(val, val, ...)
  Exemple : `v<-c(1, 7, 3, 9)` donnera `1 7 3 9`
  Exemple : `v<-c("age", "sexe")` donnera `"age" "sexe"`

Caractère / Variable :

- quantitatif
  - discret : Ne peut prendre que certaines valeurs précises.
  - continu : Peut prendre toutes les valeurs d'un inteval.
- qualitatif
- simple ou univarié : La mesure ne produit qu'une seule valeur.
- multiple : La mesure produit une série de valeurs.

Récupérer les mesures d'un caractère dans R : On va utiliser `echantillon$caractère` ce qui va renvoyer un vecteur.

Echelles quantitatives :
- Echelle ordinale : Une notion d'ordre existe entre les valeurs mesurées.
- Echelle de rapport : La notion de 0 a un sens physique (l'absence du caractère observé).
- Echelle d'intervalles : Le 0 n'a pas de sens physique. Permet uniquement la comparaison d'intervalles.

Echelle qualitative :
- Echelle nominale

Cas d'un caractère discret

Série statistique / Distribution observée / Distribution empirique :
Soit n la taille de notre échantillon S, et un caractère discret X. La suite finie notée :

$$\underline{X}(S) = (X_1, ..., X_n)$$

où X_i prend ses valeurs dans l'ensemble $\{x_1, ..., x_p\}$, avec $x_1 < x_2 < ... < x_p$.
Ou soit,

$$\underline{X}(S) = (X_i, n_i)_{i=1,...,p}$$

où n_i est le nombre de fois que la valeur x_i a été observé dans notre échantillon.

Ces séries correspondent aux valeurs atteintes par les n différents individus appartenant à notre échantillon.

Créer un tableau dans R : On va utiliser `table(vecteur)` ce qui va retourner un tableau avec les différentes valeurs (x_i) et le nombre de fois où elles sont apparues (n_i/effectif).

Effectif de la valeur x_i noté n_i : On parle d'effectif de la valeur x_i pour désigner ce nombre n_i.

Effectif cumulé N_i en x_i : Il est défini comme $\boldsymbol{N_i = \sum_{j=1}^{i} n_j}$ pour $i = 1, ..., p$.
L'effectif cumulé en x_p est tel que $\boldsymbol{x_p = n}$.

Calculer l'effectif cumulé dans R : On va utiliser `cumsum(tableau)` ce qui va renvoyer un tableau avec les effectifs cumulés.

Fréquence de la valeur x_i notée f_i : C'est le rapport de l'effectif de la valeur x_i avec l'effectif n de l'échantillon. $\boldsymbol{f_i = \frac{n_i}{n}}$

Calculer la fréquence dans R : On va utiliser `prop.table(tableau)` avec un tableau qui contient les effectifs, ce qui va renvoyer un tableau avec les fréquences.

Fréquence cumulée F_i en x_i : Elle se définit comme $\boldsymbol{F_i = \sum_{j=1}^{i} f_j}$ pour $i = 1, ..., p$.
La fréquence cumulée en x_p est tel que $\boldsymbol{f_p = 1}$.

Calculer la fréquence cumulée dans R : On va utiliser `cumsum(tableau)` avec un tableau qui contient les fréquences, ce qui va renvoyer un tableau avec les fréquences cumulées.

Cas d'un caractère continu

Règle de Sturges : Elle propose

— de déterminer le nombre de classes d'une découpe de la manière suivante :

$$\boldsymbol{nombre\ de\ classes = [log_2(n) + 1]}$$

où n est la taille de la série statistique.

— de converser une longueur de classe constante.

Calculer un regroupement en classe par la règle de Sturges dans R : On va utiliser hist(tableau) avec un tableau qui contient les effectifs, ce qui va renvoyer un histogramme.
Dans cet histogramme, on a :
— $breaks : Un vecteur qui contient les limites des classes obtenues.
— $counts : Un vecteur qui contient pour chaque classe, son effectif.
— $density : Un vecteur à ne pas confondre avec les fréquences des classes.

Effectif de $]a_i, a_{i+1}]$: C'est le nombre n_i de valeurs observées dans l'intervalle $\boldsymbol{]a_i, a_{i+1}]}$.

Calculer l'effectif dans R : On va utiliser histo$counts avec *histo* qui est un objet histogramme.

Effectif cumulé en a_i : C'est le nombre de valeurs observées dans l'intervalle $\boldsymbol{]-\infty, a_i]}$.

Calculer l'effectif cumulé dans R : On va utiliser cumsum(histo$counts) avec *histo* qui est un objet histogramme.

Distribution statistique groupée : Elle est notée :

$$\boldsymbol{(]a_i, a_{i+1}], n_i)_{i=1,...,p}}$$

avec $a_i < a_{i+1}$ et p étant le nombre de classes formants la partition de l'invervalle des valeurs possibles.

Fréquence de $]a_i, a_{i+1}]$ **notée** f_i : Elle est égale au rapport suivant : $\boldsymbol{f_i = \frac{n_i}{n}}$.

Calculer la fréquence dans R : On va utiliser histo$counts/sum(histo$counts) avec *histo* qui est un objet histogramme.

Fréquence cumulée en a_i : Elle est égale à la somme suivante : $\boldsymbol{\sum_{j=1}^{i} f_j}$ pour $i = 1, ..., p$.

Calculer la fréquence cumulée dans R : On va utiliser
cumsum(histo$counts)/sum(histo$counts) avec *histo* qui est un objet histogramme.

Cas d'un caractère qualitatif

Tableau de contingence : Il regroupe les effectifs des différentes modalités de la variable.

Créer un tableau de contingence dans R : On va utiliser table(vecteur) ce qui va retourner un tableau de contingence avec les différentes valeurs (x_i) et le nombre de fois où elles sont apparues (n_i/effectif).

Représentations graphiques

Diagramme en bâtons des effectifs (respectivement des fréquences) d'une série statistique discrète est tel que :

— en abscisse, on condidère les différentes valeurs possibles x_i, pour $i = 1, ..., p$.
— en ordonnée, on indique l'effectif (respectivement la fréquence) observé.

Créer un diagramme en bâtons dans R : On va utiliser `barplot(x, xlab?=, ylab?=)` ou `plot(x, type="h", xlab?=, ylab?=)`.

Distribution empirique d'une variable : C'est une fonction en escalier dont l'équation est la suivante :

$$\forall x \in \mathbb{R},\ F(x) = \begin{cases} 0 & \text{si } x < x_1 \\ \sum_{j=1}^{i} f_i & \text{si } x_i \leq x < x_{i+1},\ i = 1, ..., p-1 \\ 1 & \text{si } x \geq x_p \end{cases}$$

Créer une distribution empirique dans R : On va utiliser `ecdf(x)`.

Histogramme : C'est une représentation graphique d'une série statistique groupée où une barre (ou rectangle) est associée pour chaque classe. La surface de ce rectangle est proportionnelle à l'effectif de la dite classe.
On distingue 2 cas :

— Lorsque les amplitudes des classes sont égales, la base du rectangle est égale à cette amplitude et la hauteur est proportionnelle avec le même paramètre K à l'effectif de chaque classe.
— Lorsque les amplitudes sont différentes, il existe un commun diviseur a. Dès lors, la base de chaque rectangle sera proportionnelle à l'amplitude de la classe mais sera un multiple entier de ce diviseur a. La hauteur est proportionnelle avec le paramètre K à l'effectif divisé par le rapport de l'amplitude avec le diviseur a.

Créer un histogramme dans R : On va utiliser `hist(x, xlab?=, ylab?=)`.

La boîte à moustache (où boxplot en anglais) : C'est un graphique où sont représentés des caractéristiques de position et de dispersion. Les valeurs extrêmes de la boîte à moustache ne représentent pas nécessairement les extrêmes des observations mais plûtot les valeurs suivantes.
borne inf $= max(Q_{0.25}(\underline{X}) - 1.5(Q_{0.75}(\underline{X}) - Q_{0.25}(\underline{X})),\ min(\underline{X}))$
borne inf $= min(Q_{0.75}(\underline{X}) + 1.5(Q_{0.75}(\underline{X}) - Q_{0.25}(\underline{X})),\ max(\underline{X}))$

Créer une boîte à moustache dans R : On va utiliser `boxplot(x, xlab?=, ylab?=)`

Caractéristiques numériques

Soit l'échantillon discret suivant $\underline{X} = (X_i)_{i=1,...,n}$ où X_i est l'observation réalisée sur l'individu i, sachant que l'échantillon compte n individus.

Différentes représentations des observations :
— En utilisant les effectifs : Soit $\underline{X} = (x_i, n_i)_{i=1,...,p}$ avec $x_i < x_{i+1}$, pour tout $i < p$.
— En classant ces observations par ordre croissant nommée *la statistique d'ordre* : Soit la suite $(\underline{X}) = (X_{(i)})_{i=1,...,n}$.

Caractéristiques de position

Mode de la série statistique discrète $(\underline{X}) = (X_{(i)})_{i=1,...,n}$ **noté** $Mo(\underline{X})$: C'est la valeur x_i dont la fréquence est maximale.

Calculer le mode dans R : On va utiliser `which(tableau==max(tableau)`.

Différentes distributions :
- Distributions unimodales : Avec un seul mode.
- Distributions plurimodales : Avec plusieurs modes.

Quantile d'ordre α noté $Q_\alpha(\underline{X})$: Pour une série statistique discrète $(\underline{X}) = (X_{(i)})_{i=1,\dots,n}$ où $\alpha \in]0,\ 1[$, il est tel que :

$$Q_\alpha(\underline{X}) = X_{(m)} + d(X_{(m+1)} - X_{(m)})$$

où

$$m = \lfloor \alpha(n+1) \rfloor$$
$$d = \alpha(n+1) - m$$

Le premier quartile : Il est noté $Q_{0.25}(\underline{X})$.

La second quartile (ou médiane) : Il est noté $Q_{0.5}(\underline{X})$.

Le troisième quartile : Il est noté $Q_{0.75}(\underline{X})$.

Calculer un quantile dans R : On va utiliser `quantile(vecteur, vecteur_proba)`.
Exemple : quantile(tableau, c(0.25, 0.5, 0.75)) pour afficher $Q_{0.25}$, $Q_{0.5}$, $Q_{0.75}$.

Moyenne arithmétique notée $\overline{X}$: Soit l'échantillon discret $\underline{X} = (x_i, n_i)_{i=1,\dots,p}$, la moyenne arithmétique de cet échantillon est donnée par :

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{p} n_i x_i$$

Si nous travaillons avec la représentation $\underline{X} = (X_i)_{i=1,\dots,n}$, la moyenne se calcule de la manière suivante :

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} n_i x_i$$

Calculer le moyenne dans R : On va utiliser `mean(vecteur)`.

Caractéristiques de dispersion

L'étendue d'un échantillon $\underline{X}$: C'est la différence entre la plus grande et la plus petite valeur de l'échantillon, soit :

$$e(\underline{X}) = max(\underline{X}) - min(\underline{X})$$

Calculer l'étendue dans R : On va utiliser `diff(range(vecteur))`.

L'étendue interquartile notée EIQ($\underline{X}$) : C'est la différence :

$$EIQ(\underline{X}) = Q_{0.75}(\underline{X}) - Q_{0.25}(\underline{X})$$

Calculer l'étendue interquartile dans R : On va utiliser `IQR(vecteur)`.

La variance empirique notée $S^2(\underline{X})$: Elle se calcule de la manière suivante :

$$S^2(\underline{X}) = \frac{1}{n}\sum_{i=1}^{p}(x_i - \overline{X})^2 n_i$$

La variance empirique corrigée notée $S_c^2(\underline{X})$: Elle se calcule de la manière suivante :

$$S_c^2(\underline{X}) = \frac{1}{n-1}\sum_{i=1}^{p}(x_i - \overline{X})^2 n_i$$

Calculer la variance empirique corrigée dans R : On va utiliser `var(vecteur)`.

L'écart-type empirique noté $S(\underline{X})$: Il se calcule de la manière suivante :

$$S(\underline{X}) = \sqrt{S^2(\underline{X})}$$

L'écart-type empirique corrigé noté $S_c(\underline{X})$: Il se calcule de la manière suivante :

$$S_c(\underline{X}) = \sqrt{S_c^2(\underline{X})}$$

Calculer l'écart-type empirique corrigé dans R : On va utiliser `sd(vecteur)`.

Le moment centré d'ordre r noté $m_r(\underline{X})$: Pour une série statistique discrète $\underline{X} = (x_i, n_i)$ pour $i = 1, ..., p$, il est défini comme :

$$m_r(\underline{X}) = \frac{1}{n}\sum_{i=1}^{p}(x_i - \overline{X})^r n_i$$

Le moment d'ordre r noté $\overline{X^r}$: Il se calcule de la manière suivante :

$$\overline{X^r} = \frac{1}{n}\sum_{i=1}^{p} x_i^r n_i$$

Formule de Huygens : La variance empirique respecte l'identité suivante :

$$S^2(\underline{X}) = \overline{X^2} - \overline{X}^2$$

où $\overline{X}^2$ est le moment d'ordre 2 de $\underline{X}$.

Le coefficient de variation d'un échantillon $\underline{X}$ noté $CV(\underline{X})$: Il est défini par le rapport entre l'écart-type empirique et la moyenne arithmétique. Soit :

$$CV(\underline{X}) = \frac{S(\underline{X})}{\overline{X}}$$

Caractéristiques de forme

Le coefficient d'asymétrie de Fisher d'une série statistique $\underline{X}$ noté $\gamma_1(\underline{X})$: Il se calcule de la manière suivante :

$$\gamma_1(\underline{X}) = \frac{m_3(\underline{X})}{S^3(\underline{X})}$$

Calculer le coefficient d'asymétrie de Fisher dans R : On va utiliser `skewness(x)` en utilisant le package *moments*.

Le coefficient d'asymétrie de Pearson d'une série statistique $\underline{X}$ noté $\beta_1(\underline{X})$: Il est défini comme le carré du coefficient d'asymétrie de Fisher. Soit :

$$\beta_1(\underline{X}) = \gamma_1^2(\underline{X})$$

Calculer le coefficient d'asymétrie de Pearson dans R : On va utiliser `kurtosis(x)` en utilisant le package *moments*.

Le coefficient d'aplatissement de Fisher d'une série statistique $\underline{X}$ noté $\gamma_2(\underline{X})$: Il se calcule de la manière suivante :

$$\gamma_2(\underline{X}) = \frac{m_4(\underline{X})}{m_2^2(\underline{X})} - 3$$

Le coefficient d'aplatissement de Pearson d'une série statistique $\underline{X}$ noté $\beta_2(\underline{X})$: Il se calcule de la manière suivante :

$$\beta_2(\underline{X}) = \frac{m_4(\underline{X})}{S^4(\underline{X})}$$

Chapitre 2 : Statistiques descriptive bivariée

Conditions d'analyse :

Soit $(\underline{X}, \underline{Y})$ la distribution statistique d'un couple de variables mesurées sur un échantillon dont l'effectif total est n. On a plusieurs manières de décrire cette distribution, à savoir

— sous forme de données brutes, soit (X_i, Y_i) pour $i = 1, ..., n$,

— sous forme de valeurs distinctes. Soit $(x_1, ..., x_p)$ avec $x_1 < x_2 ... < x_p$ pour la variable X et de la même manière pour la variable Y ayant q valeurs, nous avons alors le *tableau de contigence*

	y_1	...	y_j	...	y_q
x_1	n_{11}	...	n_{1j}	...	n_{1q}
$\vdots$					
x_i	n_{i1}	...	n_{ij}	...	n_{iq}
$\vdots$					
x_p	n_{p1}	...	n_{pj}	...	n_{pq}

où n_{ij} représente l'*effectif* de l'échantillon pour lequel sont observées la mesure x_i pour la variable X et la mesure y_j pour la variable Y, avec $i \in \{1, \dots, p\}$ et $j \in \{1, \dots, q\}$.

Créer un data frame dans R : On va utiliser `data.frame(v1, v2, ..., vj, row.names?=)` ce qui va retourner le data frame avec les vecteurs (v_j) encodés dedans. Notons que ces vecteurs doivent être de la même taille. Par défaut, les lignes sont numérotées mais on peut choisir leurs nom en donnant un vecteur à l'argument row.names.

Le tableau de contigence fournit ce qu'on appelle la *distribution jointe* de la série statistique $(\underline{X}, \underline{Y})$. On peut aisément y adjoindre les totaux des lignes et des colonnes. On obtient alors la représentation suivante

	y_1	...	y_j	...	y_q	Total
x_1	n_{11}	...	n_{1j}	...	n_{1q}	$n_{1\bullet}$
$\vdots$						
x_i	n_{i1}	...	n_{ij}	...	n_{iq}	$n_{i\bullet}$
$\vdots$						
x_p	n_{p1}	...	n_{pj}	...	n_{pq}	$n_{p\bullet}$
Total	$n_{\bullet 1}$	...	$n_{\bullet j}$	...	$n_{\bullet q}$	n

Notions statistiques avec deux variables :

Dès lors, à l'aide de la représentation précédente, on peut définir les notions suivantes

— **l'effectif marginal** de x_i est donné par

$$n_{i\bullet} \overset{\text{def}}{=} \sum_{j=1}^{q} n_{ij},$$

celui de y_j sera défini comme

$$n_{\bullet j} \overset{\text{def}}{=} \sum_{j=1}^{p} n_{ij},$$

Obtenir l'effectif marginal dans R : On va utiliser `cumsum(data$column)` avec data notre data frame et column une colonne y_j. Cela aura peut de sens d'effectuer cette opération pour une ligne x_i. On utilisera également les commandes vu au chapitre 1 pour toutes les commande suivantes au détail près qu'on bloque une ligne/colonne.

— **la fréquence marginale** de de x_i est donnée par

$$f_{i\bullet} \overset{\text{def}}{=} \sum_{j=1}^{q} \frac{n_{ij}}{n},$$

celle de y_j sera définie comme

$$f_{\bullet j} \overset{\text{def}}{=} \sum_{i=1}^{p} \frac{n_{ij}}{n},$$

— **la distribution marginale** des fréquences de X est $(x_i, f_{i\bullet})_{i=1,...,p}$, celle de Y est $(y_i, f_{\bullet i})_{i=1,...,q}$, avec

$$\begin{aligned} f_{i\bullet} &\overset{\text{def}}{=} \frac{n_{i\bullet}}{n}, \\ f_{\bullet i} &\overset{\text{def}}{=} \frac{n_{\bullet i}}{n}. \end{aligned}$$

La fréquence conditionnelle de la valeur y_j sachant que $X = x_i$, notée $f_{j|i}^{Y|X}$ est donnée par

$$\begin{aligned} f_{j|i}^{Y|X} &\overset{\text{def}}{=} \frac{n_{ij}}{n_{i\bullet}} \\ &= \frac{f_{ij}}{f_{i\bullet}} \end{aligned}$$

La fréquence conditionnelle de la valeur x_i sachant que $Y = y_j$, notée $f_{i|j}^{X|Y}$ est donnée par

$$\begin{aligned} f_{i|j}^{X|Y} &\overset{\text{def}}{=} \frac{n_{ij}}{n_{\bullet j}} \\ &= \frac{f_{ij}}{f_{\bullet j}} \end{aligned}$$

La moyenne conditionnelle de la variable X sachant que $Y = y_j$ est definie par

$$\begin{aligned} \overline{X}_j &\overset{\text{def}}{=} \frac{1}{n_{\bullet j}} \sum_{i=1}^{p} n_{ij} x_i \\ &= \sum_{i=1}^{p} f_{i|j}^{X|Y} x_i. \end{aligned}$$

La moyenne conditionnelle de la variable Y sachant que $X = x_i$ est definie par

$$\begin{aligned} \overline{Y}_i &\overset{\text{def}}{=} \frac{1}{n_{i\bullet}} \sum_{j=1}^{q} n_{ij} y_j \\ &= \sum_{j=1}^{q} f_{j|i}^{Y|X} y_j. \end{aligned}$$

La variance conditionnelle de la variable X sachant que $Y = y_j$ est définie pas

$$\begin{aligned} S_j^2(\underline{X}) &\overset{\text{def}}{=} \frac{1}{n_{\bullet j}} \sum_{i=1}^{p} n_{ij} (x_i - \overline{X_j})^2 \\ &= \sum_{i=1}^{p} f_{i|j}^{X|Y} x_i^2 - \overline{X_j}^2. \end{aligned}$$

La variance conditionnelle de la variable Y sachant que $X = x_i$ est définie pas

$$\begin{aligned} S_i^2(\underline{Y}) &\overset{\text{def}}{=} \frac{1}{n_{i\bullet}} \sum_{j=1}^{q} n_{ij} (y_j - \overline{Y_i})^2 \\ &= \sum_{j=1}^{q} f_{j|i}^{Y|X} y_j^2 - \overline{Y_i}^2. \end{aligned}$$

Notion d'indépendance :

Les séries statistiques $\underline{X}$ et $\underline{Y}$ sont dites indépendantes si et seulement si pour tout $i \in \{1, ..., p\}$ et pour tout $j \in \{1, ..., q\}$, nous avons

$$\begin{aligned} f_{i|j}^{X|Y} &= f_{i\bullet} \\ f_{j|i}^{Y|X} &= f_{\bullet j}. \end{aligned}$$

Dès lors, on a que si $\underline{X}$ et $\underline{Y}$ sont indépendantes alors

$$f_{ij} = f_{i\bullet} f_{\bullet j}$$

pour tout $i \in \{1, ..., p\}$ et pour tout $j \in \{1, ..., q\}$. **La covariance** entre $\underline{X}$ et $\underline{Y}$, notée $\text{Cov}(\underline{X}, \underline{Y})$ est définie de la manière suivante

$$Cov(\underline{X}, \underline{Y}) \overset{\text{def}}{=} \sum_{1 \leq i \leq p, 1 \leq j \leq q} (x_i - \overline{X})(y_j - \overline{Y}) f_{ij}$$

ou de manière équivalente

$$Cov(\underline{X}, \underline{Y}) \overset{\text{def}}{=} \sum_{1 \leq i \leq p, 1 \leq j \leq q} x_i y_j f_{ij} - \overline{XY}.$$

Obtenir la covariance dans R : On va utiliser $\boxed{\text{Cov}(x, y)}$ avec x et y des vecteur de données.

Pour la covariance, on observe les propriétés suivantes

$$\begin{aligned} Cov(\underline{X}, \underline{Y}) &= Cov(\underline{Y}, \underline{X}) \\ Cov(\underline{X}, \underline{X}) &= S^2(\underline{X}) \\ |Cov(\underline{X}, \underline{Y})| &\leq \sqrt{S^2(\underline{X}) S^2(\underline{Y})} \end{aligned}$$

Le coefficient de corrélation entre $\underline{X}$ et $\underline{Y}$ noté $R(\underline{X}, \underline{Y})$ est défini de la manière suivante

$$R(\underline{X}, \underline{Y}) \overset{\text{def}}{=} \frac{Cov(\underline{X}, \underline{Y})}{S(\underline{X}) S(\underline{Y})}.$$

Le coefficient de corrélation est tel que

$$|R(\underline{X}, \underline{Y})| \leq 1.$$

Obtenir le coefficient de corrélation dans R : On va utiliser $\boxed{\text{Cor}(x, y)}$ avec x et y des vecteur de données. On peut également préciser la méthode utilisée à du paramètre 'method', comme par exemple la méthode de Spearman.

Soit la série statistique $(\underline{X}, \underline{Y})$ avec le même tableau de contigence que précédemment utilisé. Soit

$$n^*_{jk} \overset{\text{def}}{=} \frac{n_{j\bullet} n_{\bullet k}}{n}.$$

On peut mesurer l'écart à l'indépendance comme

$$e_{jk} \overset{\text{def}}{=} n_{jk} - n^*_{jk}.$$

La mesure du Khi-deux, notée D^2, donne la mesure de l'association qui existe entre deux variables nominales. Elle est donnée par

$$D^2 \overset{\text{def}}{=} \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{e^2_{jk}}{n^*_{jk}}.$$

Soit la série statistique $(\underline{X}, \underline{Y}) = (X_i, Y_i)$ pour $i = 1, ..., n$. Soit $R(x)$ le rang attribué à la valeur x. **Le coefficient de corrélation de rang de Spearman**, noté r_s est défini comme

$$r_S \overset{def}{=} \frac{1/n \sum_{i=1}^{n} (R(X_i) - \overline{R}_X)(R(Y_i) - \overline{R}_Y)}{\sqrt{(1/n \sum_{i=1}^{n} [R(X_i) - \overline{R}_X]^2)(1/n \sum_{i=1}^{n} [R(Y_i) - \overline{R}_Y]^2)}}$$

où

$$\begin{aligned} \overline{R}_X &\overset{\text{def}}{=} 1/n \sum_{i=1}^{n} R(X_i) \\ &= \frac{n+1}{2}. \end{aligned}$$

Le coefficient de corrélation de rang de Spearman respecte aussi l'égalité suivante

$$r_s = \frac{6 \sum_{i=1}^{n} (R(X_i) - R(Y_i))^2}{n(n^2-1)}$$

et

$$-1 \leq r_s \leq 1.$$

La régression linéaire :

La différence entre Y et son interprétation linéaire $aX + b$ peut s'exprimer par la quantité suivante

$$d(a,b) \overset{\text{def}}{=} \sum_{i=1}^{n} (Y_i - aX_i - b)^2$$

On dit que **la droite de régression linéaire** de Y par rapport à X est

$$Y = aX + b$$

Elle est définie de manière à minimiser la quantité $d(a,b)$. On a donc que

$$\begin{aligned} a &= \frac{Cov(\underline{X}, \underline{Y})}{S^2(\underline{X})} \\ b &= \overline{Y} - a\overline{X} \end{aligned}$$

où

- $\overline{X}$ (respectivement $\overline{Y}$) est la moyenne arithmétique de la série statistique $\underline{X}$ (respectivement de la série $\underline{Y}$),
- $\text{Cov}(\underline{X}, \underline{Y})$ est la covariance entre X et Y,
- et $S^2(\underline{X})$ est la variance empirique de X.

Obtenir la droite de régression dans R : On va utiliser `lm()`

Chapitre 3 : Analyse combinatoire

Principe fondamental :

Supposons
- qu'on réalise r expériences, numérotées de 1 à r,
- que le résultat de chaque exprérience n'influence pas le résultat des autres expériences,
- que l'expérience n°$i (i \in \{1, ..., r\})$ produise l'un des n_i résultats, alors il y aura un total de

$$n_1 n_2 ... n_r$$

résultats pour les r expériences prises ensemble.

Notion d'arrangement :

Supposons qu'on réalise une expérience où il s'agit d'arranger r objets distincts parmi n objets particuliers (avec $n > r$), et cela sans remise, alors il y aura un total de

$$n.(n-1).(n-2)...(n-r+1) = \frac{n!}{(n-r)!}$$

arrangements possibles.

Supposons le même cas de figure mais avec des remises. Alors il y aura un total de

$$n^r$$

arrangements possibles.

Notion de permutation :

Supposons qu'on réalise une expérience où il s'agit de permuter r objets particuliers distincts les uns des autres, alors il y aura un total de

$$r.(r-1)...1 = r!$$

résultats possibles pour cette expérience.

Supposons qu'on réalise une expérience où il s'agit de permuter ; objets particuliers dont n_1 sont indiscernables entre eux,..., n_r sont indiscernables entre eux, alors il y aura un total de

$$\frac{n!}{n_1! n_2! ... n_r!}$$

résultats possibles pour cette expérience. On parle alors de **permutations multiples**.

Notion de combinaison :

Supposons qu'on réalise une expérience où r objets doivent être choisis parmi n objets distincts où l'ordre d'apparition des objets n'est pas significatif, alors on aura

$$\binom{n}{r} \underset{=}{\text{def}} \frac{n!}{(n-r)!r!}$$

combinaisons possibles.

Supposons qu'on réalise une expérience où r objets doivent être classifiés parmi n sous-ensembles pouvant contenir chacun jusqu'à r objets, où l'ordre d'appartenance des objets n'est pas significatif, alors on aura

$$\binom{n+r-1}{r} \underset{=}{\text{def}} \frac{(n+r-1)!}{(n-1)!r!}$$

façons possibles.

Chapitre 4 : Le calcul des probabilités

Ensemble fondamental

Ensemble fondamental noté Ω : C'est l'ensemble des valeurs possibles pour une expérience.

Un événement : C'est un sous-ensemble de l'ensemble fondamental Ω. Si le résultat d'une expérience est connu et est une des valeurs reprises dans l'événement E, on dit que l'événement E est *réalisé*.

Propriétés : Soit E et F deux événement d'un ensemble fondamental Ω.

- On note $E \cup F$ l'événement qui contient les éléments appartenant à E ou à F. Ainsi l'événement $E \cup F$ est réalisé si et seulement si l'événement E ou l'événement F est réalisé.
- On note $E \cap F$ l'événement qui contient les éléments appartenant à E ou à F. Ainsi l'événement $E \cap F$ est réalisé si et seulement si l'événement E et l'événement F est réalisé.
- Les ensembles E et F sont *mutuellement exclusifs* lorsque $E \cap F = \emptyset$

Axiomes

Probabilités de réalistion d'un événement : Pour définir cette notion, nous partons de 3 axiomes. Dans un premier temps, pour chaque événement E, il existe une valeur appelée *probabilité*.

La probabilité notée P(E) : C'est une valeur qui satisfait les trois axiomes suivants :

$$\begin{aligned} 0 \leq P(E) \leq 1 \\ P(\Omega) = 1 \\ P(\bigcup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} P(E_i) \end{aligned}$$

lorsque les événement E_i, poir $i = 1, ..., \infty$ sont mutuellement exclusifs.

$\overline{\boldsymbol{E}}$: C'est l'événement complémentaire à E dans Ω (ensemble fondamental).

Déduction :

- $P(\emptyset) = 0$
- $P(\overline{E}) = 1 - P(E)$

Théorème :

- Soit deux événement E et F définis sur Ω et tels que $E \subseteq D$, alors $P(E) \leq P(F)$.
- Soit deux événement E et F définis sur Ω, alors $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Probabilités sur des ensembles finis

Supposons que l'ensemble Ω soit un ensemble d'éléments discrets, de taille finie, soit un ensemble fini.
Probabilité d'un événement E noté P(E) : Cela revient à compter le nombre d'éléments en faisant partie soit :

$$P(E) = \frac{\#E}{N}$$

avec N rappelons le, égal au cardinal de Ω, soit $\#\Omega$.

N'importe quelle mesure (tant qu'elle respecte les trois axiomes) peut être utilisée pour donner la probabilité qu'un événement se réalise.

Probabilités conditionnelle

Considérons les deux événements E et F définis sur l'ensemble fondamental Ω et P(F) > 0.
La probabilité conditionnelle de *E* étant donné la réalisation de *F* notée P(E|F) : C'est la probabilité que E se réalise sachant que F se réalise et elle se calcule comme :

$$P(E|F) = \frac{P(E \cap F)}{P(F)}$$

Dans le contexte d'une mesure conditionnelle, nous devons nous placer dans l'ensemble des éléments de F. Ces éléments définissent notre "nouvel espace fondamental".

De cet ensemble, nous retenons les éléments qui permettent également la réalisation de E. Ces éléments constituent un ensemble, soit $E \cap F$, dont la probabilité est $P(E \cap F)$.

Règle de multiplication

Soit les événements $E_1, ..., E_n$ définis sur l'ensemble fondamental Ω.
La règle de la multiplication : Elle s'écrit comme :

$$P(E_1 \cap E_2 \cap ... \cap E_n) = P(E_1)P(E_2|E_1)P(E_3|E_1 \cap E_2)...P(E_n|E_1 \cap E_2 \cap ... \cap E_{n-1})$$

Formule des probabilités totales

La formule des probabilités totales : Elle nous permet de calculer P(E) de la manière suivante :

$$P(E) = P(E|F)P(F) + P(E|\overline{F})(1 - P(F))$$

Soit les événements F_i avec $i = 1, ..., n$ s'excluant mutuellement et tels que :

$$\bigcup_{i=1}^{n} F_i = \Omega$$

La formule des probabilités totales généralisée : Elle se calcule de la manière suivante :

$$\begin{aligned} P(E) &= \sum_{i=1}^{n} P(E \cap F_i) \\ &= \sum_{i=1}^{n} P(E|F_i)P(F_i) \end{aligned}$$

Formule de Bayes

Soit E et F deux événements définis sur Ω.
La formule de Bayes : Elle se calcule de la manière suivante :

$$P(E|F) = \frac{P(F|E)P(E)}{P(F|E)P(E) + P(F|\overline{E})P(\overline{E})}$$

Soit les événements F_i avec $i = 1, ..., n$ s'excluant mutuellement et tels que $\bigcup_{i=1}^{n} F_i = \Omega$.
La formule de Bayes généralisée : Elle se calcule de la manière suivante :

$$P(F_j|E) = \frac{P(E|F_j)P(F_j)}{\sum_{i=1}^{n} P(E|F_i)P(F_i)}$$

Distribution conditionnelle

Les probabilités conditionnelles satisfont aux trois axiomes des probabilités.

Axiomes : Soit E et F deux événements issus de l'ensemble fondamental Ω, tels que $P(F) \neq 0$.

1. $0 \leq P(E|F) \leq 1$
2. $P(\Omega|F) = 1$
3. Soit E_i, pour $i = 1, ..., n$, des événements qui s'excluent mutuellement. Alors :

$$P(\bigcup_{i=1}^{n} E_i|F) = \sum_{i=1}^{n} P(E_i|F)$$

Indépendance

Un événement indépendant : Soit deux événements E et F issus tous les deux de l'ensemble fondamental Ω. On dit que l'événement E est indépendant de l'événement F lorsque :

$$P(E \cap F) = P(E)P(F)$$

Lorsque deux événements sont indépendants, nous avons :

$$P(E|F) = P(E)$$

Soit deux événements E et F issus tous deux de l'ensemble fondamental Ω. Supposons que E et F sont indépendant. Dans ce cas, E et $\overline{F}$ le sont aussi.

Evénements totalement indépendants

Des événements totalement indépendants :

— Soit E, F et G trois événements issus de l'ensemble fondamental Ω. Ces événements sont dit totalement indépendants lorsque les conditions suivantes sont satisfaites :

$$\begin{aligned} P(E \cap F \cap G) &= P(E)P(F)P(G) \\ P(E \cap F) &= P(E)P(F) \\ P(F \cap G) &= P(F)P(G) \\ P(E \cap G) &= P(E)P(G) \end{aligned}$$

— Ou soit, les événements $E_1, E_2, ..., E_n$ sont dit totalement indépendants lorsque pour tout sous-ensemble $E_{1'}, ..., E_{r'}$ où $(1', 2', ..., r')$ est une combinaison de r éléments de l'ensemble d'indices {1, 2,...,n}, avec $r \leq n$, nous avons :

$$P(E_{1'} \cap ... \cap E_{r'}) = P(E_{1'}...P(E_{r'}))$$

Indépendance conditionnelle

Deux événements E_1 et E_2 conditionnellement indépendants selon F : Soit trois événements E_1, E_2 et F issus d'un même ensemble fondamental Ω. Deux événements E_1 et E_2 sont dit conditionnellement indépendants selon F si la probabilité conditionnelle de E_1 (E_2 respectivement) étant donné que F est réalisé, n'est pas affectée par l'occurence ou non de E_2 (E_1 respectivement), soit formellement :

$$\begin{aligned} P(E_1|E_2 \cap F) &= P(E_1|F) \\ P(E_2|E_1 \cap F) &= P(E_2|F) \end{aligned}$$

ou encore

$$P(E_1 \cap E_2|F) = P(E_1|F)P(E_2|F)$$

Chapitre 5 : Variables aléatoires

Definitions fondamentales :

Soit $X : \Omega \to \mathbb{R}$, une fonction qui prends ses valeurs dans l'ensemble fondamental Ω et à valeurs réelles. On suppose que l'ensemble fondamental Ω est un espace mesurable dont la mesure est la robabilité $P(A)$, définie pour tout événement A défini sur Ω.

La fonction X est une variable aléatoire lorsqu'elle est mesurable. Dès lors, nous avons, pour tout $x \in \mathbb{R}$

$$Pr[X \leq x] = P(\{\omega \in \Omega | X(\omega) \leq x\}).$$

Toute **fonction de répartition** F vérifie les propriétés suivantes
- la fonction F est une fonction non-décroissante.
- $$\lim_{b \to \infty} F(b) = 1,$$
- $$\lim_{b \to -\infty} F(b) = 0.$$
- La fonction F est continue à droite.

On observe la propriété suivante

$$Pr[a < X \leq b] = F(b) - F(a).$$

Variables aléatoires discrètes :

Un variable aléatoire $X : \Omega \to \mathbb{R}$ est dite discrète lorsqu'elle prend ses valeurs dans un ensemble dénombrable, inclus dans $\mathbb{R}$. La loi de probabilité de X est dans ce cas définie comme la fonction $p(.)$, avec

$$p(a) \stackrel{\text{def}}{=} Pr[X = a]$$

appelée **fonction de densité** de la variable aléatoire X.

La fonction de répartition F_X d'une variable aléatoire discrète X est obtenue à partir de la fonction de densité p_X de cette variable aléatoire de la manière suivante

$$F_X(a) = \sum_{i : i \leq a} p_X(i).$$

L'espérance de la variable aléatoire discrète X, notée $\mathbb{E}[X]$ est donnée par l'expression

$$\mathbb{E}[X] \stackrel{\text{def}}{=} \sum_{i : i > 0} ip(i),$$

où $p(.)$ est la loi de la variable aléatoire X.

Soit la fonction g définie de $\mathbb{R}$ dans $\mathbb{R}$. Soit X une variable aléatoire discrète prenant les valeurs x_i avec la probabilité $p(x_i)$ (où $p(x_i) \neq 0$), alors

$$\mathbb{E}[g(X)] = \sum_i g(x_i)p(x_i).$$

Soit X une variable aléatoire discrète et deux constantes a et b. On a

$$\mathbb{E}[aX + b] = a\mathbb{E}[X] + b.$$

La variance de la variable aléatoire discrète X, notée $\text{Var}[X]$ se calcule de la manière suivante

$$\begin{aligned} Var[X] &\stackrel{\text{def}}{=} \mathbb{E}[(X - \mathbb{E}[X])^2] \\ &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \end{aligned}$$

L'écart-type de cette variable aléatoire se définit comme

$$\sigma \stackrel{\text{def}}{=} \sqrt{Var[X]}$$

Soit X une variable aléatoire discrète et deux constantes a et b. On a

$$Var[aX + b] = a^2 Var[X].$$

Soit X une variable aléatoire discrète de loi $p(.)$; le moment d'ordre k noté $\mathbb{E}[X^k]$ est défini comme

$$\mathbb{E}[X^k] \stackrel{\text{def}}{=} \sum_{i:i>0} i^k p(i),$$

pour $k \in \mathbb{N}$.

Le moment centré d'ordre k est quant à lui défini comme

$$\mathbb{E}[(X-\mu)^k] \stackrel{\text{def}}{=} \sum_{i:i>0} (i-\mu)^k p(i),$$

pour $k \in \mathbb{N}$ et où $\mu = \mathbb{E}[X]$.

Variables aléatoires continues :

On qualifie X variable aléatoire de **variable aléatoire continue** lorsqu'il existe une fonction f non négative définie pour tout $x \in \mathbb{R}$ telle que pour tout ensemble $B \subseteq \mathbb{R}$, on observe

$$Pr[X \in B] \stackrel{\text{def}}{=} \int_B f(x)dx,$$

où la fonction f est appelée **densité de probabilité** de la variable aléatoire X.

Les fonctions de répartition F et densité f d'une variable aléatoire X continue sont liées de la manière suivante

$$\frac{d}{da}F(a) = f(a)$$

avec a réel.

L'espérance d'une variable aléatoire X continue, de densité f et notée $\mathbb{E}[X]$, est donnée par

$$\mathbb{E}[X] \stackrel{\text{def}}{=} \int_{-\infty}^{\infty} x f(x)dx.$$

On observe les propriétés suivantes

— avec X non négative,

$$\mathbb{E}[X] \stackrel{\text{def}}{=} \int_0^{\infty} Pr[X > x] f(x)dx$$

— avec g une fonction réelle,

$$\mathbb{E}[X] \stackrel{\text{def}}{=} \int_{-\infty}^{\infty} g(x) f(x)dx$$

— avec a et b, deux constantes,

$$\mathbb{E}[aX + b] = a\mathbb{E}[X] + b.$$

Soit une variable aléatoire X continue à fonction de densité f, sa **variance**, notée Var$[X]$, se défini comme

$$\begin{aligned} Var[X] &\stackrel{\text{def}}{=} \mathbb{E}[(X - \mathbb{E}[X])^2] \\ &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2. \end{aligned}$$

L'écart-type de cette variable aléatoire se définit comme

$$\sigma \stackrel{\text{def}}{=} \sqrt{Var[X]}$$

Soit X une variable aléatoire continue et deux constantes a et b. On a

$$Var[aX + b] = a^2 Var[X].$$

On appelle quantile d'ordre q de la variable continue X, où $q \in [0, 1]$, la valeur x_q telle que $F(x_q) = q$ avec F fonction de répartition de la variable X.

Soit la variable aléatoire continue, X de fonction de densité $f(.)$; le moment d'ordre k, noté $\mathbb{E}[X^k]$, est donné par

$$\mathbb{E}[X^k] \stackrel{\text{def}}{=} \int_{-\infty}^{\infty} x^k f(x) dx,$$

pour $k \in \mathbb{N}$.

Le moment centré d'ordre k est quant à lui défini comme

$$\mathbb{E}[(X - \mu)^k] \stackrel{\text{def}}{=} \int_{-\infty}^{\infty} (x - \mu)^k f(x) dx,$$

pour $k \in \mathbb{N}$ et où $\mu = \mathbb{E}[X]$.

Fonction génératrice et transformée de Laplace :

Soit un réel z, la fonction génératrice des moments, notée $M_X(z)$, de la variable aléatoire X discrète est définie comme

$$M_X(z) \stackrel{\text{def}}{=} \mathbb{E}[z^X].$$

Soit une variable aléatoire X discrète, nous avons

$$\mathbb{E}[X(X-1)\dots(X-k+1)] = \frac{d^k}{dz^k} M_X(z)\Big|_{z=1}.$$

La fonction caractéristique de la variable aléatoire X continue, notée $\psi_X(s)$ pour s réel, est définie comme

$$\psi_X(s) \stackrel{\text{def}}{=} \mathbb{E}[\exp\{isX\}].$$

Lorsque la variable aléatoire est positive, on parle encore de transformée de Laplace, notée $\phi_X(s)$ et définie pour s complexe dont la partie réelle est positive, de la manière suivante

$$\phi_X(s) \stackrel{\text{def}}{=} \mathbb{E}[\exp\{-sX\}].$$

Soit X une variable aléatoire continue et à valeurs positives, et soit $\phi_X(.)$ sa transformée de Laplace. Alors

$$(-1)^k \mathbb{E}[X^k] = \frac{d^k}{ds^k} \phi_X(s)\Big|_{s=0}.$$

Chapitre 6 : Lois de probabilités usuelles

Lois discrètes :

Nom et cas d'utilisation	$\Pr[X]$	$\mathbb{E}[X]$ et $\mathbb{E}[X^k]$	$\mathrm{Var}[X]$
La variable aléatoire de Bernouilli permet de décrire le succès (avec la probabilité p) ou l'échec d'une expérience.	$Pr[X=1]=p$ $Pr[X=0]=1-p$	$\mathbb{E}[X]=p$	$\mathrm{Var}[X]=p(1-p)$
La variable aléatoire binomiale s'utilise lors d'une expérience où on réalise n tirages indépendants avec la probabilité p que le tirage soit un succès. La variable X représente le nombre de succès sur n tirages.	$Pr[X=i]$ $\overset{\text{def}}{=} \binom{n}{i}p^i(1-p)^{(n-i)}$	$\mathbb{E}[X]=np$ $\mathbb{E}[X^k]=np\mathbb{E}[(Y+1)^{k-1}]$ avec Y, $\mathrm{bin}(n-1,p)$	$\mathrm{Var}[X]=np(1-p)$
Variable aléatoire de Poisson. On ne donne pas de cas d'utilisation.	$Pr[X=i]$ $\overset{\text{def}}{=} \exp\{-\lambda\}\frac{\lambda^i}{i!}$	$\mathbb{E}[X]=\lambda$	$\mathrm{Var}[X]=\lambda$
La variable aléatoire géométrique permet d'observer l'instant de la première occurence d'un succès pour n tirages avec la probabilité p de succès.	$Pr[X=i]$ $\overset{\text{def}}{=} (1-p)^{i-1}p$	$\mathbb{E}[X]=\frac{1}{p}$	$\mathrm{Var}[X]=\frac{1-p}{p^2}$
La variable aléatoire binomiale négative permet de trouver n tel qu'il faut réaliser n expériences pour observer le r^{ieme} succès.	$Pr[X=n]$ $\overset{\text{def}}{=} \binom{n-1}{r-1}p^r(1-p)^{(n-r)}$	$\mathbb{E}[X]=\frac{r}{p}$	$\mathrm{Var}[X]=\frac{p(1-p)}{p^2}$

Nom et cas d'utilisation	$\Pr[X]$	$\mathbb{E}[X]$ et $\mathbb{E}[X^k]$	$\mathrm{Var}[X]$
La variable aléatoire hypergéométrique permet donne la probabilité par exemple d'observer i boules blanches sur un tirage de n boules dans une urnes contenant N boules blanches et noires dont m sont blanches.	$Pr[X = i] \underset{\text{def}}{=} \frac{\binom{m}{i}\binom{N-m}{n-i}}{\binom{N}{n}}$	$\mathbb{E}[X] = \frac{nm}{N}$	$\mathrm{Var}[X] = \frac{nm}{N}(1 - \frac{m}{N})(1 - \frac{n-1}{N-1})$
La variable aléatoire uniforme discrète est telle que si elle prend n valeurs distinctes pour chaque valeur k on a une probabilité de $\frac{1}{n}$ de l'observer.	$Pr[X = k] = \frac{1}{n}$	$\mathbb{E}[X] = \frac{n+1}{2}$	$\mathrm{Var}[X] = \frac{n^2-1}{12}$

Nom et cas d'utilisation	$\Pr[X = i] = x$	$\Pr[X \leq i] = x$	$\Pr[X \leq x] = p$
La variable aléatoire de Bernouilli.	/	/	/
La variable aléatoire binomiale.	dbinom$(i, size, proba)$	pbinom$(i, size, proba)$	qbinom$(p, size, proba)$
Variable aléatoire de Poisson.	dpois$(i, lambda)$	ppois$(i, lambda)$	qpois$(p, lambda)$
La variable aléatoire géométrique.	dgeom$(i, proba)$	pgeom$(i, proba)$	qgeom$(p, proba)$
La variable aléatoire binomiale négative.	dnbinom$(i, size, proba)$	pnbinom$(i, size, proba)$	qnbinom$(p, size, proba)$
La variable aléatoire hypergéométrique.	dhyper(x, m, n, k)	phyper(q, m, n, k)	qhyper(p, m, n, k)
La variable aléatoire uniforme discrète.			

Lois continues :

Nom et cas d'utilisation	Pr$[X]$	$\mathbb{E}[X]$ et $\mathbb{E}[X^k]$	Var$[X]$	Commande R
La variable aléatoire uniforme continue, comme sont homologue discrète, est uniformément distibuée sur l'intervalle (a, b).	$Pr[X \leq b] = \int_{-\infty}^{b} \frac{1}{b-a}$	$\mathbb{E}[X] = \frac{a+b}{2}$	Var$[X] = \frac{(b-a)^2}{12}$	/
La variable aléatoire normale.	$Pr[X \leq a] = \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}\sigma} \exp\{\frac{-(x-\mu)^2}{2\sigma^2}\}$	$\mathbb{E}[X] = \mu$	Var$[X] = \sigma^2$	Pour obtenir $Pr[X \leq a]$ on peut utiliser $pnorm(a)$. Attention elle doit être centrée réduite, autrement dit, on fait $\frac{X-\mu}{\sigma}$.Dès lors, il faut effectuer $pnorm(a\sigma + \mu)$
La distribution exponentielle est habituellement utilisée pour représenter un temps d'attente avant l'arrivée d'un événement particulier.	$Pr[X \leq x] = \lambda \exp\{-\lambda x\}$ avec $x \geq 0$ et 0 sinon.	$\mathbb{E}[X] = \frac{1}{\lambda}$	Var$[X] = \frac{1}{\lambda^2}$	/
On ne nous en dit pas plus sur la **loi d'Erlang**.	$Pr[X \leq x] = \frac{\lambda \exp\{-\lambda x\}(\lambda x)^{n-1}}{(n-1)!}$ avec $x \geq 0$ et 0 sinon.	$\mathbb{E}[X] = \frac{n}{\lambda}$	Var$[X] = \frac{n}{\lambda^2}$	/

Approximations de la binomiale :

On pourrait aproximer le loi binomiale de différentes manières.

1. Elle peut être approchée d'aussi près que l'on veut par une loi hypergéométrique de paramètre N, pN, n avec $N \to \infty$.
2. Soit S_n le nombre de succès lors de n épreuves indépendantes, la probabilité de réussite pour chaque épreuve étant p. Alors pour tout $a < b$, on peut écrire

$$Pr\left[a \leq \frac{S_n - np}{\sqrt{np(1-p)}} \leq b\right] \to \Phi(b) - \Phi(a)$$

   avec $n \to \infty$ Dès lors on à que S_n est une binomiale à laquelle on à retranché sont espérance et divisée par son écart-type. Dès cette variable se rapproche d'une loi normale pour n assez grand.

3. Enfin, on pourrait démontrer que que l'on peut également approximer la binomiale avec une loi de Poisson de paramètre np.

Fonction génératrice et transformée de Laplace :

Lois discrètes Voici un tableau reprenant la loi p(x) et la fonction génératrice M(t) des différentes lois discrètes importante.

Loi	p(x)	M(z)
$\mathrm{Bin}(n,p)$	$\binom{n}{x} p^x (1-p)^{n-x}$	$(1-p+pz)^n$
$\mathrm{Pois}(\lambda)$	$\exp\{-\lambda\}\frac{\lambda^x}{x!}$	$\exp\{-\lambda(1-z)\}$
$\mathrm{Geom}(p)$	$p\,(1-p)^{x-1}$	$\frac{1-p}{1-pz}$
$\mathrm{Bin.nég}(r,p)$	$\binom{x-1}{r-1} p^r (1-p)^{x-r}$	$\left(\frac{p}{1-qz}\right)^r$

Lois continues Voici le tableau concernant les lois continues. Lorsque la loi est positive, nous présentons la transformée de Laplace $\phi(s)$ plutôt que la fonction caractéristique $\psi(s)$.

Loi	f(x)	$\phi(s)/\psi(s)$
$U(a,b)$	$\begin{cases} \frac{1}{b-a} & a<x<b \\ 0 & \text{sinon} \end{cases}$	$\psi(s)=\frac{\exp\{sb\}-\exp\{sa\}}{s(b-a)}$
$N(\mu,\sigma^2)$	$\frac{1}{\sqrt{2\pi}\sigma}\exp\{-(x-\mu)^2/2\sigma^2\}$	$\psi(s)=\exp\{\mu s+\sigma^2 s^2/2\}$
$\exp(\lambda)$	$\begin{cases} \lambda\exp\{-\lambda x\} & x\geq 0 \\ 0 & x<0 \end{cases}$	$\phi(s)=\frac{\lambda}{\lambda+s}$
$\mathrm{Erlang}(n,\lambda)$	$\begin{cases} \frac{\lambda\exp\{-\lambda x\}(\lambda x)^{n-1}}{(n-1)!} & \text{si } x\geq 0 \\ 0 & \text{sinon.} \end{cases}$	$\phi(s)=\left(\frac{\lambda}{\lambda+s}\right)^n$

Chapitre 7 : Variables aléatoires simultanées

Distribution jointe - cas continu

Une fonction F de répartition jointe des variables aléatoires X et Y définie sur le même espace fondamental Ω : Elle est donnée par :

$$\begin{aligned} F(a,b) &= Pr[X \leq a, Y \leq b] \\ &= P(\{\omega \in \Omega \mid X(\omega) \leq a \wedge Y(\omega) \leq b\}) \end{aligned}$$

pour tout $a, b \in \mathbb{R}$.

Propriétés

— $Pr[X \leq a, Y \leq b] = Pr[X \leq a \wedge Y \leq b]$.

— La fonction de répartition jointe mesure donc la probabilité d'observer simultanément $X \leq a$ et $Y \leq b$.

La fonction f(x, y) de densité jointe de deux variables continues X, Y étudiées simulanément : Elle est définie pour tout-ensemble $A \times B \subseteq \mathbb{R}^2$ de la manière suivante :

$$Pr[X \in A, Y \in B] = \int_A \int_B f(x,y)dxdy$$

Nous avons alors :

$$Pr[X \leq A, Y \leq B] = \int_{-\infty}^{a} \int_{-\infty}^{b} f(x,y)dxdy$$

Dès lors :

$$\frac{\partial^2}{\partial x \partial y} F(x,y) = f(x,y)$$

Distribution jointe - cas discret

Une loi p de probabilité jointe des variables aléatoires X et Y discrètes définies sur le même espace fondamental Ω : Elle est donnée par :

$$p(a,b) = Pr[X = a, Y = b]$$

pour tout $a, b \in \mathbb{N}$.

Distribution jointe - n variables

La fonction de répartition jointe F de n variables $X_1, X_2, ..., X_n$: Elle se calcule de la manière suivante :

$$\begin{aligned} F(a_1, ..., a_n) &= Pr[X_1 \leq a_1, ..., X_n \leq a_n] \\ &= P(\{\omega \in \Omega \mid X_1(\omega) \leq a_1 \wedge ... \wedge X_n(\omega) \leq a_n\}) \end{aligned}$$

Soit $i = 1, ..., n$ et les variables X_i continues
La densité f jointe : Elle se définie comme :

$$Pr[X_1 \in A_1, ..., \ X_n \in A_n] = \int_{A_1} \int_{A_2} ... \int_{A_n} f(x_1, ..., x_n)dx_1dx_2...dx_n$$

Soit $i = 1, ..., n$ et les variables X_i discrètes
La loi de probabilité jointe : Elle se définie comme :

$$p(a_1, a_2, ..., a_n) = Pr[X_1 = a_1, X_2 = a_2, ..., X_n = a_n]$$

Théorèmes :

— Soit X et Y deux variables aléatoires *discrètes*, de loi jointe p(x, y), alors pour toute fonction g(x, y) :

$$\mathbb{E}(g(X,Y)) = \sum_x \sum_y g(x,y)p(x,y)$$

— Soit X et Y deux variables aléatoires *continues*, où f(x, y) est la densité jointe de X et Y, alors pour toute fonction g(x, y) :

$$\mathbb{E}(g(X,Y)) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)f(x,y)dxdy$$

Distribution marginale

La fonction de répartion marginale de X notée $F_X(.)$:

$$\begin{aligned} F_X(a) &= Pr[X \leq a] \\ &= Pr[X \leq a, Y < \infty] \\ &= F(a, \infty) \end{aligned}$$

La fonction de répartion marginale de Y notée $F_Y(.)$:

$$\begin{aligned} F_Y(b) &= Pr[Y \leq b] \\ &= Pr[X < \infty, Y \leq b] \\ &= F(\infty, b) \end{aligned}$$

Déductions :

$$\begin{aligned} Pr[X > a, Y > b] &= P(\{\omega \in \Omega \mid X(\omega) > a \wedge Y(\omega) > b\}) \\ &= 1 - P(\{\omega \in \Omega \mid X(\omega) > a \wedge Y(\omega) > b\}^c) \\ &= 1 - P(\{\omega \in \Omega \mid X(\omega) \leq a \vee Y(\omega) \leq b\}) \\ &= 1 - Pr[X \leq a \vee Y \leq b] \\ &= 1 - (Pr[X \leq a] + Pr[Y \leq b] - Pr[X \leq a, Y \leq b]) \\ &= 1 - F_X(a) - F_Y(b) + F(a,b) \end{aligned}$$

$$Pr[a_1 < X \leq a_2, b_1 < Y \leq b_2] = F(a_2,b_2) + F(a_1,b_1) - F(a_1,b_2) - F(a_2,b_1)$$

Densité marginale de X :

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy$$

Densité marginale de Y :

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx$$

Soit X et Y des variables discrètes
Loi de probabilité marginale de X

$$\begin{aligned} p_X(a) &= Pr[X = a] \\ &= \sum_{y \,:\, p(a,y)>0} p(a,y) \end{aligned}$$

Loi de probabilité marginale de Y

$$\begin{aligned} p_Y(b) &= Pr[Y = b] \\ &= \sum_{x \,:\, p(x,b)>0} p(x,b) \end{aligned}$$

Covariance

Soit deux variables aléatoires X et Y
La covariance notée Cov(X, Y) : Elle est définie par :

$$\begin{aligned} Cov(X,Y) &= \mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])] \\ &= \mathbb{E}[XY]-\mathbb{E}[X]\mathbb{E}[Y] \end{aligned}$$

Dans le cas d'une variable aléatoire *discrète* où p(x, y) est la loi de probabilité jointe de X et de Y, nous avons alors :

$$\mathbb{E}[XY] = \sum_{x,y} x\ y\ p(x,y)$$

Dans le cas d'une variable aléatoire *continue* où f(x, y) est la densité jointe de X et de Y, nous avons alors :

$$\mathbb{E}[XY] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x\ y\ f(x,y)dxdy$$

Propriétés : Soit X et Y deux variables aléatoires quelquonques, alors :
- Cov(X, Y) = Cov(Y, X)
- Cov(X, X) = Var(X)
- Cov(aX, Y) = a Cov(X, Y)

Corrélation

La corrélation entre deux variables aléatoires X et Y notée $\rho(X,Y)$: Elle est définie comme :

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

sous l'hypothèse que Var(X) et Var(Y) sont bien différentes de 0.

Variables aléatoires indépendantes

Indépendances de deux variables aléatoires : Deux variables aléatoires X et Y sont dites indépendantes lorsque pour tout sous-ensemble A et B de $\mathbb{R}$, nous avons :

$$Pr[X \in A, Y \in B] = Pr[X \in A]Pr[Y \in B]$$

Nous avons donc :

$$Pr[X \leq a, Y \leq b] = Pr[X \leq a]Pr[Y \leq b]$$

soit

$$F(a,b) = F_X(a)F_Y(b).$$

Si de plus, les variables sont continues, nous aurons

$$f(a,b) = f_X(a)f_Y(b)$$

Dans les cas de lois discrètes, nous avons enfin

$$p(a,b) = p_X(a)p_Y(b)$$

Indépendances de n variables aléatoires : Les n ($n < \infty$) variables aléatoires $X_1, X_2, ..., X_3$ sont dites indépendantes si pour toute collection de r sous-ensembles $A_{1'}, A_{2'}, ..., A_{r'}$ avec $(1', 2', ..., r')$ une combinaison de r éléments choisis parmi (1, 2,..., n), nous avons :

$$Pr[X_{1'} \in A_{1'}, ..., X_{r'} \in A_{r'}] = \prod_{i=1'}^{r'} Pr[X_i \in A_i]$$

Lorsque n est infini, la collection infinie de variables aléatoires est dite indépendante si tout sous-ensemble fini que l'on puisse en tirer est composé de variables indépendantes.

Propriétés :

— La covariance de deux variables aléatoires X et Y indépendantes est nulle.

— Lorsque la covariance de deux variables aléatoires est nulle, cela ne signifie pas nécessairement que X et Y sont indépendantes.

Somme de deux variables aléatoires indépendantes

Nous avons :

$$\begin{aligned} Pr[X+Y \leq a] &= F_{X+Y}(a) \\ &= \int_{-\infty}^{\infty} F_X(a-y) f_Y(y) dy \end{aligned}$$

Ainsi :

$$\begin{aligned} f_{X+Y}(a) &= \frac{\partial}{\partial a} \int_{-\infty}^{\infty} F_X(a-y) f_Y(y) dy \\ &= \int_{-\infty}^{\infty} f_X(a-y) f_Y(y) dy \end{aligned}$$

Propriétés :

— Soit X_i, pour $i = 1, ..., n$, n variables aléatoires quelconques, leur espérance est telle que :

$$\mathbb{E}[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} \mathbb{E}[X_i]$$

sous l'hypothèse que $\mathbb{E}[X_i]$ est finie pour tout $i \in 1, ..., n$.

— Soit X_i, pour $i = 1, ..., n$ et Y_j, pour $j = 1, ..., m$, n+m variables aléatoires quelconques, alors :

$$Cov(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j) = \sum_{i=1}^{n} \sum_{j=1}^{m} Cov(X_i, Y_i)$$

— Soit X_i, pour $i = 1, ..., n$, n variables aléatoires quelconques. Nous avons :

$$Var(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} Var(X_i) + \sum_{i=1}^{n} \sum_{j=1}^{n} Cov(X_i, X_j)$$

avec $j \neq i$.

— Soit X et Y deux variables aléatoires quelconques telles que Var(X) et Var(Y) sont bien différentes de 0, alors :

$$-1 \leq \rho(X, Y) \leq 1$$

— Soit X_i, pour $i = 1, ..., n$, n variables aléatoires indépendantes deux à deux. Nous avons :

$$Var(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} Var(X_i)$$

La suite X_i, pour $i = 1, ..., n$, de n variables aléatoires quelconques, est composée de variables aléatoires indépendantes deux à deux lorsque les variables aléatoires X_i et X_j pour tout i, j différents, sont indépendantes.

Sommes remarquables

Théorèmes :

— Soit X_1, X_2,...,X_n, n variables aléatoires indépendantes normales de paramètre (μ_i, σ_i^2), pour $i = 1, 2, ..., n$. La somme $\sum_{i=1}^{n} X_i$ est une variable aléatoire de *distribution normale* de paramètre $(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2)$.

— Soit X_1, X_2,...,X_n, n variables aléatoires indépendantes de Poisson de paramètre λ_i, $i = 1, 2, ..., n$. La somme $\sum_{i=1}^{n} X_i$ est une variable aléatoire de *distribution de Poisson* de paramètre $\sum_{i=1}^{n} \lambda_i$.

— Soit X_1, X_2,...,X_n, n variables aléatoires indépendantes binomiales de paramètre $(n_i$, p), i=1,2,...,n. La somme $\sum_{i=1}^{n} X_i$ est une variable aléatoire de *distribution binomiale* de paramètre $(\sum_{i=1}^{n} n_i$, p).

— Soit deux variables aléatoires indépendantes X_1 et X_2 de distribution exponentielle de paramètre λ. La somme de ces deux variables aléatoires est égale à :

$$f_{X_1+X_2}(a) = \lambda exp\{-\lambda a\}(\lambda a)$$

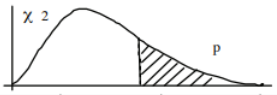
qui est la densité d'une *loi Erlang* où n = 2.

Distribution chi-carrée

Soit X_1, X_2,...,X_n, n variables aléatoires normales centrées réduites, indépendantes.
Une variable aléatoire de distribution chi-carrée à n degrés de liberté notée Z : Elle est définie comme :

$$Z = \sum_{i=1}^{n} X_i^2$$

TABLE DU CHI-DEUX : $\chi^2(n)$



n p	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.02	0.01
1	0,0158	0,0642	0,148	0,455	1,074	1,642	2,706	3,841	5,412	6,635
2	0,211	0,446	0,713	1,386	2,408	3,219	4,605	5,991	7,824	9,210
3	0,584	1,005	1,424	2,366	3,665	4,642	6,251	7,815	9,837	11,341
4	1,064	1,649	2,195	3,357	4,878	5,989	7,779	9,488	11,668	13,277
5	1,610	2,343	3,000	4,351	6,064	7,289	9,236	11,070	13,388	15,086
6	2,204	3,070	3,828	5,348	7,231	8,558	10,645	12,592	15,033	16,812
7	2,833	3,822	4,671	6,346	8,383	9,803	12,017	14,067	16,622	18,475
8	3,490	4,594	5,527	7,344	9,524	11,030	13,362	15,507	18,168	20,090
9	4,168	5,380	6,393	8,343	10,656	12,242	14,684	16,919	19,679	21,666
10	4,865	6,179	7,267	9,342	11,781	13,442	15,987	18,307	21,161	23,209
11	5,578	6,989	8,148	10,341	12,899	14,631	17,275	19,675	22,618	24,725
12	6,304	7,807	9,034	11,340	14,011	15,812	18,549	21,026	24,054	26,217
13	7,042	8,634	9,926	12,340	15,119	16,985	19,812	22,362	25,472	27,688
14	7,790	9,467	10,821	13,339	16,222	18,151	21,064	23,685	26,873	29,141
15	8,547	10,307	11,721	14,339	17,322	19,311	22,307	24,996	28,259	30,578
16	9,312	11,152	12,624	15,338	18,418	20,465	23,542	26,296	29,633	32,000
17	10,085	12,002	13,531	16,338	19,511	21,615	24,769	27,587	30,995	33,409
18	10,865	12,857	14,440	17,338	20,601	22,760	25,989	28,869	32,346	34,805
19	11,651	13,716	15,352	18,338	21,689	23,900	27,204	30,144	33,687	36,191
20	12,443	14,578	16,266	19,337	22,775	25,038	28,412	31,410	35,020	37,566
21	13,240	15,445	17,182	20,337	23,858	26,171	29,615	32,671	36,343	38,932
22	14,041	16,314	18,101	21,337	24,939	27,301	30,813	33,924	37,659	40,289
23	14,848	17,187	19,021	22,337	26,018	28,429	32,007	35,172	38,968	41,638
24	15,659	18,062	19,943	23,337	27,096	29,553	33,196	36,415	40,270	42,980
25	16,473	18,940	20,867	24,337	28,172	30,675	34,382	37,652	41,566	44,314
26	17,292	19,820	21,792	25,336	29,246	31,795	35,563	38,885	42,856	45,642
27	18,114	20,703	22,719	26,336	30,319	32,912	36,741	40,113	44,140	46,963
28	18,939	21,588	23,647	27,336	31,391	34,027	37,916	41,337	45,419	48,278
29	19,768	22,475	24,577	28,336	32,461	35,139	39,087	42,557	46,693	49,588
30	20,599	23,364	25,508	29,336	33,530	36,250	40,256	43,773	47,962	50,892

Pour n > 30, on peut admettre que $\sqrt{2\chi^2} - \sqrt{2n-1} \approx N(0,1)$

Cette table donne non pas les quantiles d'ordre p mais les quantiles d'ordre 1 - p.

Théorèmes :

— Soit X une variable aléatoire qui suit la loi chi-carrée avec p degrés de libertés. Son espérance et sa variance sont données par :

$$\mathbb{E}[X] = p$$
$$Var[X] = 2p$$

— Soit deux variables aléatoires indépendantes X_1 et X_2 de loi chi-carrée avec respectivement p_1 et p_2 degrés de libertés. Alors $X_1 + X_2$ suit une loi chi-carrée avec $p_1 + p_2$ degrés de liberté.

Distribution de Student

La loi de Student à n degrés de liberté notée $T =_d t(n)$: La variable aléatoire T suit la loi de Student à n degrés de liberté lorsque T est une variable aléatoire continue qui admet pour densité :

$$f(t) = \frac{1}{\sqrt{n\pi}} \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} \frac{1}{(1 + \frac{t^2}{n})^{\frac{n+1}{2}}}$$

pour $t \in \mathbb{R}$, avec

$$\Gamma(r) = \int_0^\infty t^{r-1} exp\{-t\} dt$$

On peut démontrer que la variable aléatoire T est obtenue par

$$T = \frac{U}{\sqrt{\frac{X}{n}}}$$

avec U variable aléatoire normale centrée réduite et X variable aléatoire de loi chi-carrée avec n degrés de libertés. Les variables U et X sont *indépendantes.*

Théorème :

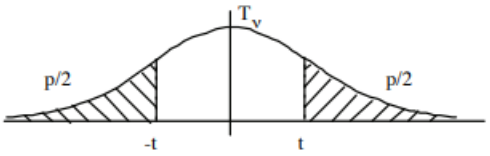
— Soit X une variable aléatoire de Student avec n degrés de libertés. Lorsque $n \geq 2$,

$$\mathbb{E}[X] = 0$$

et lorsque $n \geq 3$,

$$Var[X] = \frac{n}{n-2}$$

Variable de STUDENT à ν degrés de liberté



$T_\nu = \frac{U}{\sqrt{Y/\nu}}$ où $U \approx N(0,1)$ et $Y \approx \chi^2(\nu)$ sont indépendants en probabilité.

TABLE de t en fonction du degré de liberté ν et de la probabilité p, tels que $P(|T_\nu| > t) = p$:

ν \ p	0,90	0,70	0,50	0,40	0,30	0,20	0,10	0,05	0,02	0,01
1	0,158	0,510	1,000	1,376	1,963	3,078	6,314	12,706	31,821	63,657
2	0,142	0,445	0,816	1,061	1,386	1,886	2,920	4,303	6,965	9,925
3	0,137	0,424	0,765	0,978	1,250	1,638	2,353	3,182	4,541	5,841
4	0,134	0,414	0,741	0,941	1,190	1,533	2,132	2,776	3,747	4,604
5	0,132	0,408	0,727	0,920	1,156	1,476	2,015	2,571	3,365	4,032
6	0,131	0,404	0,718	0,906	1,134	1,440	1,943	2,447	3,143	3,707
7	0,130	0,402	0,711	0,896	1,119	1,415	1,895	2,365	2,998	3,499
8	0,130	0,399	0,706	0,889	1,108	1,397	1,860	2,306	2,896	3,355
9	0,129	0,398	0,703	0,883	1,100	1,383	1,833	2,262	2,821	3,250
10	0,129	0,397	0,700	0,879	1,093	1,372	1,812	2,228	2,764	3,169
11	0,129	0,396	0,697	0,876	1,088	1,363	1,796	2,201	2,718	3,106
12	0,128	0,395	0,695	0,873	1,083	1,356	1,782	2,179	2,681	3,055
13	0,128	0,394	0,694	0,870	1,079	1,350	1,771	2,160	2,650	3,012
14	0,128	0,393	0,692	0,868	1,076	1,345	1,761	2,145	2,624	2,977
15	0,128	0,393	0,691	0,866	1,074	1,341	1,753	2,131	2,602	2,947
16	0,128	0,392	0,690	0,865	1,071	1,337	1,746	2,120	2,583	2,921
17	0,128	0,392	0,689	0,863	1,069	1,333	1,740	2,110	2,567	2,898
18	0,127	0,392	0,688	0,862	1,067	1,330	1,734	2,101	2,552	2,878
19	0,127	0,391	0,688	0,861	1,066	1,328	1,729	2,093	2,539	2,861
20	0,127	0,391	0,687	0,860	1,064	1,325	1,725	2,086	2,528	2,845
21	0,127	0,391	0,686	0,859	1,063	1,323	1,721	2,080	2,518	2,831
22	0,127	0,390	0,686	0,858	1,061	1,321	1,717	2,074	2,508	2,819
23	0,127	0,390	0,685	0,858	1,060	1,319	1,714	2,069	2,500	2,807
24	0,127	0,390	0,685	0,857	1,059	1,318	1,711	2,064	2,492	2,797
25	0,127	0,390	0,684	0,856	1,058	1,316	1,708	2,060	2,485	2,787
26	0,127	0,390	0,684	0,856	1,058	1,315	1,706	2,056	2,479	2,779
27	0,127	0,389	0,684	0,855	1,057	1,314	1,703	2,052	2,473	2,771
28	0,127	0,389	0,683	0,855	1,056	1,313	1,701	2,048	2,467	2,763
29	0,127	0,389	0,683	0,854	1,055	1,311	1,699	2,045	2,462	2,756
30	0,127	0,389	0,683	0,854	1,055	1,310	1,697	2,042	2,457	2,750
∞	0,12566	0,38532	0,67449	0,84162	1,03643	1,28155	1,64485	1,95996	2,32634	2,57582

La loi limite, lorsque ν tend vers l'infini, est une loi Normale centrée réduite.

Loi de Fisher-Snedecor

Soit U et V deux variables aléatoires indépendantes telles que $U =_d \chi^2_{\nu_1}$ et $V =_d \chi^2_{\nu_2}$.
La loi de Fisher-Snedecor à ν_1 degrés de liberté au numérateur et ν_2 degrés de liberté au dénominateur notée $\mathbf{F}(\nu_1, \nu_2)$: Elle se définit comme :

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

Propriété :
Soit $H =_d F(\nu_1, \nu_2)$, alors

$$\frac{1}{H} =_d F(\nu_2, \nu_1)$$

Distributions conditionnelles - cas discret

Soit X et Y deux variables aléatoires discrètes.
La loi de probabilité de X sous la condition Y = y : Elle est donnée par :

$$\begin{aligned} p_{X|Y}(x|y) &= Pr[X = x|Y = y] \\ &= \frac{Pr[X = x, Y = y]}{Pr[Y = y]} \\ &= \frac{p(x, y)}{p_Y(y)} \end{aligned}$$

pour autant que $p_Y(y) > 0$.

La fonction de répartition conditionnelle de X sachant Y = y : Elle est donnée par :

$$\begin{aligned} F_{X|Y}(x|y) &= Pr[X \leq x|Y = y] \\ &= \sum_{a \leq x} p_{X|Y}(a|y) \end{aligned}$$

Clairement, lorsque X et Y sont indépendantes, nous avons :

$$p_{X|Y}(x|y) = Pr[X = x]$$

Distributions conditionnelles - cas continu

Soit X et Y deux variables aléatoires continues.
La densité conditionnelle de X sous la condition Y = y : Lorsque $f_Y(y) > 0$, elle est donnée par :

$$f_{X|Y}(x|y) = \frac{f(x, y)}{f_Y(y)}$$

Ainsi

$$Pr[X \in A|Y = y] = \int_A f_{X|Y}(x|y)dx$$

La fonction de répartition conditionnelle de X sous la condition Y = y : Elle est donnée par :

$$\begin{aligned} F_{X|Y}(a|y) &= Pr[X \leq a|Y = y] \\ &= \int_{-\infty}^{a} f_{X|Y}(x|y)dx \end{aligned}$$

Sans mémoire

Une variable aléatoire non-négative sans mémoire : Une variable aléatoire X non-négative est dite sans mémoire lorsque :

$$Pr[X > s + t|X > t] = Pr[X > s]$$

pour $s, t \geq 0$.

La classe *des variables exponentielles* est la seule classe de variables aléatoire continues à posséder la propriété d'être *sans mémoire*

Théorèmes limites

Théorèmes :

— **Inégalité de Markov** : Soit X une variable aléatoire à valeurs non-négatives. Pour tout réel a > 0, nous avons :

$$Pr[X \geq a] \leq \frac{\mathbb{E}[X]}{a}$$

— **Inégalité de Tchebychev** : Soit X une variable aléatoire d'espérance μ et de variance σ^2, alors pour tout k > 0, nous avons

$$Pr[|X - \mu| \geq k] \leq \frac{\sigma^2}{k^2}$$

— **Loi faible des grands nombres** : Soit X_i, $i \geq 1$ une suite de variables aléatoires indépendantes et identiquement distribuées. Leur espérance commune est finie et notée μ. Dès lors, pour tout $\epsilon > 0$, nous avons :

$$Pr[|\frac{\sum_{i=1}^{n} X_i}{n} - \mu| > \epsilon] \to 0$$

lorsque $n \to \infty$.

La loi faible des grands nombres nous indique que *probablement*, la variable aléatoire $\sum_{i=1}^{n} X_i/n$ resterait proche de la valeur constante μ.

— **La loi forte des grands nombres** : Soit $X_1, X_2, ...$ une suite de variables aléatoires indépendantes et identiquement distribuées d'espérance μ, avec $\mu < \infty$. Avec probabilité 1, nous avons :

$$\frac{X_1 + X_2 + ... + X_n}{n} \to \mu$$

lorsque $n \to \infty$.

Ce résultat nous indique que la moyenne arithmétique de variables aléatoires tend vers l'espérance commune de celles-ci.

— **Théorème central limite** : Soit $X_1, X_2, ...$ une suite de variable aléatoires indépendantes et identiquement distribuées d'espérance μ et de variance σ^2, alors :

$$Pr[\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} \leq a] \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} exp\{-\frac{x^2}{2}\}dx$$

lorsque $n \to \infty$.

Ce théorème nous indique sous quelles conditions la distribution de la somme d'une suite de variables aléatoires converge vers la distribution normale.

Chapitre 8 : Théorie de l'estimation

Estimation ponctuelle :

Un estimateur $\hat{\theta}_n$ de θ obtenu sur base de l'échantillon $(X_1, X_1, ..., X_n)$ est une fonction h_n de ce même échantillon. Une estimation ponctuelle du paramètre θ, est obtenue par $h_n(x_1, x_2, ..., x_n)$.

Soit un estimateur $\hat{\theta}_n$ du paramètre θ. **Le biais de l'estimateur** $\hat{\theta}_n$, noté $B(\hat{\theta}_n)$ se définit par

$$B(\hat{\theta}_n) \stackrel{\text{def}}{=} \mathbb{E}[\hat{\theta}_n] - \theta.$$

On dit qu'un est **estimateur est sans biais** quand

$$B(\hat{\theta}_n) = 0.$$

Un estimateur est dit asymptotiquement sans biais lorsque

$$\lim_{n\to\infty} \mathbb{E}[\hat{\theta}_n] = \theta.$$

Soit $\hat{\mu}_n$ défini comme

$$\hat{\mu}_n \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} X_i,$$

soit la moyenne arithmétique de l'échantillon $(X_1, X_2, \ldots, X_n)$. **L'estimateur** $\hat{\mu}_n$ **est un estimateur sans biais de** μ**, espérance théorique** de la loi X commune à tous les X_i de l'échantillon.

Le moment empirique d'ordre k $(k > 0)$, noté $\hat{m}_n^k$ est sans biais pour l'estimation du moment théorique d'ordre k de la loi X, loi parente d'un échantillon X de taille n.

L'écart quadratique moyen d'un estimateur $\hat{\theta}_n$ de θ, noté $ECQ(\hat{\theta}_n)$ est donné par

$$\begin{aligned} ECQ(\hat{\theta}_n) &\stackrel{\text{def}}{=} \mathbb{E}[(\hat{\theta}_n - \theta)^2] \\ &= Var(\hat{\theta}_n) + B(\hat{\theta}_n)^2. \end{aligned}$$

Un estimateur $\hat{\theta}_n^1$ est relativement plus efficace qu'un estimateur $\hat{\theta}_n^2$ lorsque

$$ECQ(\hat{\theta}_n^1) \leq ECQ(\hat{\theta}_n^2).$$

On dit alors que l'estimateur $\hat{\theta}_n^1$ domine l'estimateur $\hat{\theta}_n^2$. Dès lors un estimateur sans biais est optimal comparé aux autres estimateurs sans biais lorsque il est l'estimateur le plus efficace parmi tous ces estimateurs sans biais.
Enfin, **un estimateur est dit convergent** lorsqu'il est sans biais et que sa variance tend vers zéro quand n tend vers l'infini.
L'estimateur de l'espérance d'une loi parente X de variance σ^2 est donné par

$$Var[\hat{\mu}_n] = \frac{\sigma^2}{n}.$$

On observe également que

$$\begin{aligned} ECQ(\hat{\mu}_n) &= Var[\hat{\mu}_n] \\ &= \frac{\sigma^2}{n}. \end{aligned}$$

L'estimateur du paramètre p d'une loi de Bernouilli Soit un échantillon $X = (X_1, X_2, \ldots, X_n)$. Pour estimer p, on peut tout naturellement choisir la proportion empirique, soit l'estimateur $\hat{T}_1$ suivant

$$\hat{T}_1 \stackrel{\text{def}}{=} \frac{S_n}{n},$$

où n est la taille de l'échantillon et

$$S_n = \sum_{i=1}^{n} X_i.$$

Cependant, l'estimateur $\hat{T}_2$ suivant

$$\hat{T}_2 \stackrel{\text{def}}{=} \frac{S_n + 1}{n+2},$$

bien que biaisé, est préférable lorsque p est au voisinage de $1/2$. En effet, nous avons pour chaque estimateur

$$\mathbb{E}[\hat{T}_1] = p \qquad \text{Var}[\hat{T}_1] = \tfrac{p(1-p)}{n}$$

$$\mathbb{E}[\hat{T}_2] = \tfrac{np+1}{n+2} \quad \text{Var}[\hat{T}_2] = \tfrac{np(1-p)}{(n+2)^2}$$

Ainsi leur ECQ sont

$$\begin{aligned} ECQ(\hat{T}_1) &= \frac{p(1-p)}{n} \\ ECQ(\hat{T}_2) &= \frac{(1-2p)^2 + np(1-p)}{(n+2)^2}. \end{aligned}$$

Dès lors on a que quand $p = 1/2$, $\hat{T}_2$ domine $\hat{T}_1$.

L'estimateur $\hat{S}_n^2$ de la variance σ^2 d'une population défini comme

$$\hat{S}_n^2 \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} (X_i - \hat{\mu}_n)^2$$

est biaisé et son biais vaut

$$B(\hat{S}_n^2) = \frac{-\sigma^2}{n}$$

Cependant, cet estimateur est de manière asymptotique sans biais.

L'estimateur corrigé $\hat{S}_{n,c}^2$ de la variance empirique, défini comme

$$\begin{aligned} \hat{S}_{n,c}^2 &\stackrel{\text{def}}{=} \frac{nS_n^2}{n-1} \\ &= \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \hat{\mu}_n)^2 \end{aligned}$$

est un estimateur sans biais de la variance σ^2. Lorsque $X_1, X_2, \ldots, X_n$ sont des variables aléatoires normales indépendantes et identiquement distribuées de moyenne μ et de variance σ^2, alors

1. $\hat{\mu}_n$ est une variable aléatoire normale de moyenne μ et de variance σ^2/n,
2. $(n-1)\hat{S}_{n,c}^2/\sigma^2$ est une variable aléatoire chi-carrée à $n-1$ degrés de liberté.

Estimation par intervalle de confiance :

Soit $\underline{X} = (X_1, X_2, \ldots, X_n)$ un échantillon de taille n, toute fonction $h(\underline{X})$ des n variables aléatoires X_i est une statistique.
Une procédure d'intervalle de confiance de niveau α pour le paramètre θ est un couple de statistiques $(T_1(\underline{X}), T_2(\underline{X}))$ calculée sur l'échantillon $\underline{X}$ tel que

$$Pr[T_1(\underline{X}) \leq \theta \leq T_2(\underline{X})] \geq \alpha.$$

L'intervalle de confiance de niveau α, noté $IC_\alpha(\theta)$ est obtenu en considérant un échantillon $x = (x_1, x_2, \ldots, x_n)$ soit

$$IC_\alpha(\theta) = [T_1(\underline{X}), T_2(\underline{X})].$$

Estimation de l'espérance d'une loi normale :

Méthode 8.1 :
Hypothèses :
- Soit $\underline{x}$ un échantillon réalisé, de taille n, de loi parente X normale de paramètre (μ, σ^2).
- La variance σ^2 est connue.

Intervalle de confiance :
L'intervalle de confiance pour l'espérance μ au niveau de confiance α est égal à

$$\left[\hat{\mu}_n(\underline{x}) - z_{(1+\alpha)/2}\frac{\sigma}{\sqrt{n}}, \hat{\mu}_n(\underline{x}) - z_{(1-\alpha)/2}\frac{\sigma}{\sqrt{n}}\right],$$

où
- $\hat{\mu}_n(\underline{x})$ est l'estimateur de l'espérance calculé sur base de l'observation $\underline{x} = (x_1, x_2, \ldots, x_n)$. Cet estimateur est la moyenne arithmétique,
- z_α est le quantile d'ordre α de la loi normale centrée réduite.

Méthode 8.2 :
Hypothèses :
- Soit $\underline{x}$ un échantillon réalisé, de taille n, de loi parente X normale de paramètre (μ, σ^2).
- La variance σ^2 est inconnue.

Intervalle de confiance :
L'intervalle de confiance pour l'espérance μ au niveau de confiance α est égal à

$$\left[\hat{\mu}_n(\underline{x}) - t_{n-1;(1+\alpha)/2}\frac{\hat{S}_{n,c}(\underline{x})}{\sqrt{n}}, \hat{\mu}_n(\underline{x}) - t_{n-1;(1-\alpha)/2}\frac{\hat{S}_{n,c}(\underline{x})}{\sqrt{n}}\right],$$

où
- $\hat{\mu}_n(\underline{x})$ est l'estimateur de l'espérance calculé sur base de l'observation $\underline{x} = (x_1, x_2, \ldots, x_n)$. Cet estimateur est la moyenne arithmétique,
- $\hat{S}_{n,c}$ est la racine carrée de la variance empirique corrigée calculée sur base de l'observation $x = (x_1, x_2, ..., x_n)$,
- $t_{n-1;\alpha}$ est le quantile d'ordre α de la loi Student à $(n-1)$ degrés de liberté.

Estimation de la variance d'une loi normale :

Méthode 8.3 :
Hypothèses :
- Soit $\underline{x}$ un échantillon réalisé, de taille n, de loi parente X normale de paramètre (μ, σ^2).
- L'espérance μ est inconnue.

Intervalle de confiance :
L'intervalle de confiance pour la variance σ^2 au niveau de confiance α est égal à

$$\left[\frac{(n-1)\hat{S}_{n,c}(\underline{x})}{\chi^2_{n-1;(1+\alpha)/2}}, \frac{(n-1)\hat{S}_{n,c}(\underline{x})}{\chi^2_{n-1;(1-\alpha)/2}}\right],$$

où
- $\hat{S}_{n,c}$ est la racine carrée de la variance empirique corrigée calculée sur base de l'observation $x = (x_1, x_2, ..., x_n)$,
- $\chi^2_{n-1;\alpha}$ est le quantile d'ordre α de la loi chi-carrée à $(n-1)$ degrés de liberté.

Estimation d'une probabilité ou d'une proportion :

Méthode 8.4 :
Hypothèses :
Soit $\underline{x}$ un échantillon réalisé, de taille n et de loi parente commune X où π_A donne la proportion d'individus dans la population présentant le caractère A. **Intervalle de confiance :**
L'intervalle de confiance pour la proportion π_A au niveau de confiance α est égal à

$$\left[\hat{\pi}_{n,A}(\underline{x}) + z_{(\alpha-1)/2}\sqrt{\frac{\hat{\pi}_{n,A}(\underline{x})(1-\hat{\pi}_{n,A}(\underline{x}))}{n}}, \hat{\pi}_{n,A}(\underline{x}) + z_{(\alpha+1)/2}\sqrt{\frac{\hat{\pi}_{n,A}(\underline{x})(1-\hat{\pi}_{n,A}(\underline{x}))}{n}}\right],$$

où

- $\hat{\pi}_{n,A}(\underline{x})$ est l'estimateur de la proportion calculé sur base de l'observation $\underline{x} = (x_1, x_2, ..., x_n)$. Il s'agit simplement de la proportion d'individus portant le caractère A.
- z_α est le quantile d'ordre α de la loi normale centrée réduite.

Estimation du paramètre λ d'une variable aléatoire de Poisson :

Méthode 8.5 :
Hypothèses :

- Soit $\underline{x}$ un échantillon réalisé, de taille n et de loi parente commune X de loi de Poisson de paramètre λ.
- $n \geq 30$.

Intervalle de confiance :
L'intervalle de confiance pour le paramètre λ au niveau de confiance α est égal à

$$\left[\hat{\mu}_n(\underline{x}) + z_{(1-\alpha)/2}\sqrt{\frac{\hat{\mu}_n(\underline{x})}{n}}, \hat{\mu}_n(\underline{x}) + z_{(1+\alpha)/2}\sqrt{\frac{\hat{\mu}_n(\underline{x})}{n}}\right],$$

où

- $\hat{\mu}_n(\underline{x})$ est la moyenne arithmétique calculée sur base de l'observation $\underline{x} = (x_1, x_2, ..., x_n)$.
- z_α est le quantile d'ordre α de la loi normale centrée réduite.

Chapitre 9 : Tests d'hypothèse

Définitions générales

On oppose deux hypothèses complémentaire H_0 et H_1
- — $\boldsymbol{H_0}$: C'est l'hypothèse que l'on aimerait pouvoir *rejeter*.
- — $\boldsymbol{H_1}$: C'est l'hypothèse que l'on aimerait pouvoir *accepter*.

La raison est que mathématiquement, on peut prendre une décision en contrôlant l'erreur de première espèce, soit l'erreur de refuser H_0 alors qu'il aurait été plus avisé de la conserver.

Un test statistique : C'est une procédure de décision, réalisée avec un risque de première espèce contrôlé.

Un test statistique significatif : Un test est *significatif* dès lors que l'on peut réfuter l'hypothèse nulle H_0 avec un risque contrôlé.

Lorsqu'on ne peut pas rejeter l'hypothèse nulle, on dit par convention, qu'on ne peut rien conclure même si H_0 semble plausible. Dans ce cas, en effet, on ne contrôle plus le risque de prendre une mauvaise décision.

Différents tests :
- — **bilatéral** : Lorsque H_1 est de la forme $H_1 : \theta \neq ...$
- — **unilatéral** : Lorsque H_1 est de la forme $H_1 : \theta < ...$ ou $H_1 : \theta > ...$

9.1 Test de l'espérance d'une population normale de variance connue

Conditions : Soit $\underline{\mathrm{X}} = (H_1, H_2, ..., H_n)$ un échantillon aléatoire simple, issu d'une population $\mathrm{N}(\mu, \sigma^2)$, avec σ^2 connue et soit $\underline{\mathrm{x}}$ un échantillon réalisé.

Hypothèses : Nous réalisons un test bilatéral. Nous désirons donc choisir entre :

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$

Statistique de test :

$$T(\underline{X}) = \frac{\hat{\mu}_n - \mu_0}{\sigma/\sqrt{n}}$$

Loi sous H_0 : Nous savons que lorsque l'hypothèse H_0 est vérifiée, alors :

$$T(\underline{X}) =_d N(0,1)$$

Règle de comportement : Il faut *rejeter l'hypothèse nulle*, pour un risque α, lorsque la statistique $\mathrm{T}(\underline{\mathrm{x}})$ calculée sur l'échantillon réalisé $\underline{\mathrm{x}}$, soit :

$$T(\underline{x}) = \frac{\hat{\mu}_n(\underline{x}) - \mu_0}{\sigma/\sqrt{n}}$$

est telle que

$$T(\underline{x}) \in [\mu_0 - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \mu_0 + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}]$$

où z_u est le quantile d'ordre u pour une distribution normale centrée réduite.

9.2 Test de l'espérance d'une population normale de variance inconnue

Conditions : Soit $\underline{\mathrm{X}} = (X_1, X_2, ..., X_n)$ un échantillon aléatoire simple, issu d'une population $N(\mu, \sigma^2)$, avec σ^2 inconnue et $\underline{\mathrm{x}}$ un échantillon réalisé.

Hypothèses : Nous réalisons un test *bilatéral*. Nous désirons donc choisir entre

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$

Statistique de test :

$$T(\underline{X}) = \frac{\hat{\mu}_n - \mu_0}{\hat{S}_{n,c}(\underline{X})/\sqrt{n}}$$

Loi sous H_0 : Nous savons que sous l'hypothèse nulle, $T(\underline{X}) =_d$ loi de Student avec $(n-1)$ degrés de liberté.

Règle de comportement : Il faut *rejeter l'hypothèse nulle*, pour un risque α lorsque la statistique $T(\underline{x})$, calculée pour l'échantillon réalisé $\underline{x}$, soit :

$$T(\underline{X}) = \frac{\hat{\mu}_n(\underline{x}) - \mu_0}{\hat{S}_{n,c}(\underline{x})/\sqrt{n}}$$

est telle que

$$T(\underline{x}) \notin [-t_{n-1,1-\alpha/2}, t_{n-1,1-\alpha/2}]$$

ou lorsque

$$\hat{\mu}_n(\underline{x}) \notin [\mu_0 - t_{n-1,1-\alpha/2}\frac{\hat{S}_{n,c}(\underline{x})}{\sqrt{n}}, \mu_0 + t_{n-1,1-\alpha/2}\frac{\hat{S}_{n,c}(\underline{x})}{\sqrt{n}}]$$

où $t_{n,u}$ est le quantile d'ordre u pour une distribution student à n degrés de liberté.

Test de l'espérance d'une population normale de variance inconnue dans R : On va utiliser `t.test(vecteur, alternative=, mu=, conf.level?=)`.

9.3 Test de l'espérance d'une population quelconque de variance connue

Conditions : Soit $\underline{X} = (X_1, X_2, ..., X_n)$ un échantillon aléatoire simple, issu d'une population de loi quelconque, avec σ^2 connue et soit $\underline{x}$ un échantillon réalisé.

Hypothèses : Nous réalisons un test bilatéral. Nous désirons donc choisir entre :

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$

Statistique de test :

$$T(\underline{X}) = \frac{\hat{\mu}_n - \mu_0}{\sigma/\sqrt{n}}$$

Loi sous H_0 : Par le théorème central limite, on peut démontrer que :

$$T(\underline{X}) =_d N(0,1)$$

Règle de comportement : Il faut *rejeter l'hypothèse nulle*, avec un risque α lorsque la statistique de test calculée sur l'échantillon réalisé $\underline{x}$, soit :

$$T(\underline{x}) = \frac{\hat{\mu}_n(\underline{x}) - \mu_0}{\sigma/\sqrt{n}}$$

est telle que

$$T(\underline{x}) \notin \left[-z_{1-\alpha/2}, z_{1-\alpha/2}\right]$$

où z_u est le quantile d'ordre u pour une distribution normale centrée réduite.

9.4 Test de l'espérance d'une population quelconque de variance inconnue

Conditions : Soit $\underline{X} = (X_1, X_2, ..., X_n)$ un échantillon aléatoire simple, issu d'une population de loi quelconque, avec σ^2 inconnue et soit $\underline{x}$ un échantillon réalisé.

Hypothèses : Nous réalisons un *test bilatéral*. Nous désirons donc choisir entre :

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$

Statistique de test :

$$T(\underline{X}) = \frac{\hat{\mu}_n - \mu_0}{\hat{S}_{n,c}(\underline{X})/\sqrt{n-1}}$$

Loi sous H_0 : On peut démontrer que :

$$T(\underline{X}) =_d N(0,1)$$

Règle de comportement : Il faut *rejeter l'hypothèse nulle*, avec un risque α lorsque la statistique T($\underline{x}$) calculée sur l'échantillon réalisé $\underline{x}$, soit :

$$T(\underline{X}) = \frac{\hat{\mu}_n(\underline{x}) - \mu_0}{\hat{S}_{n,c}(\underline{X})/\sqrt{n-1}}$$

est telle que

$$T(\underline{x}) \notin [-z_{1-\alpha/2}, z_{1-\alpha/2}]$$

où z_u est le quantile d'ordre u pour une distribution normale centrée réduite.

9.5 Test de comparaison des espérances issues de deux populations de loi normale de variance inconnue

Conditions : Soit

- $\underline{X} = (X_1, X_2, ..., X_{n_1})$ un échantillon aléatoire simple, issu d'une population de N(μ_1, σ_1^2), avec σ_1^2 inconnue.
- $\underline{Y} = (Y_1, Y_2, ..., Y_{n_2})$ un échantillon aléatoire simple, issu d'une population de N(μ_2, σ_2^2), avec σ_2^2 inconnue.

On suppose également que ces deux échantillons sont *indépendants entre eux.*
Soit $\underline{x}$ et $\underline{y}$ deux échantillons réalisés correspondants.

Hypothèses : Nous réalisons un test *bilatéral*. Nous désirons donc choisir entre :

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

Statistique de test :

$$T(\underline{X}, \underline{Y}) = \frac{(\hat{\mu}_{X,n_1} - \hat{\mu}_{Y,n_2})}{\sqrt{\frac{\hat{S}^2_{X,c}}{n_1-1} + \frac{\hat{S}^2_{Y,c}}{n_2-1}}}$$

où $\hat{\mu}_{X,n_1}$ (respectivement $\hat{\mu}_{Y,n_1}$) est la moyenne arithmétique pour l'échantillon $\underline{X}$ ($\underline{Y}$ respectivement) et $\underline{S}^2_{X,c}$ (respectivement $\underline{S}^2_{Y,c}$) est la variance corrigée pour l'échantillon $\underline{X}$ ($\underline{Y}$ respectivement).

Loi sous H_0 : On peut démontrer que :

$T(\underline{X}, \underline{Y}) =_d$ Student à ν degrés de liberté

avec ν l'entier le plus proche de :

$$\frac{(\frac{\hat{S}^2_{X,c}}{n_1-1}+\frac{\hat{S}^2_{Y,c}}{n_2-1})^2}{\frac{\hat{S}^4_{X,c}}{(n_1-1)n_1^2}+\frac{\hat{S}^4_{Y,c}}{(n_2-1)n_2^2}}$$

Règle de comportement : Il faut *rejeter l'hypothèse nulle* avec un risque α lorsque la statistique de test calculée sur les échantillons réalisés, soit :

$$T(\underline{x},\underline{y}) = \frac{(\hat{\mu}_{X,n_1}(\underline{x}) - \hat{\mu}_{Y,n_2}(\underline{y}))}{\sqrt{\frac{\hat{S}^2_{X,c}}{n_1-1}+\frac{\hat{S}^2_{Y,c}}{n_2-1}}}$$

est telle que

$$T(\underline{x},\underline{y}) \notin [-t_{\nu,1-\alpha/2}, t_{\nu,1-\alpha/2}]$$

où $t_{n,u}$ est le quantile d'ordre u pour une distribution Student à n degrés de liberté.

Test de comparaison des espérances issues de deux populations de loi normale de variance inconnue dans R : On va utiliser
`t.test(vecteur1, vecteur2, paired=FALSE, alternative=, var.equal=FALSE, conf.level?=)`.

9.6 Test de la variance d'une population normale d'espérance connue

Conditions : Soit $\underline{X} = (X_1, X_2, ..., X_n)$ un échantillon aléatoire simple, issu d'une population $N(\mu, \sigma^2)$, avec μ connue et soit un échantillon réalisé $\underline{x}$.

Hypothèses : Nous réalisons un test *bilatéral*. Nous désirons donc choisir entre :

$$H_0 : \sigma = \sigma_0$$
$$H_1 : \sigma \neq \sigma_0$$

Statistique de test :

$$T(\underline{X}) = \sum_{i=1}^{n}(\frac{X_i-\mu}{\sigma_0})^2$$

Loi sous H_0 : On peut démontrer que

$$T(\underline{X}) =_d X^2_n$$

Règle de comportement : Il faut *rejeter l'hypothèse nulle*, avec un risque α lorsque la statistique calculée sur l'échantillon réalisé $\underline{x}$, soit :

$$T(\underline{x}) = \sum_{i=1}^{n}(\frac{x_i-\mu}{\sigma_0})^2$$

est telle que

$$T(\underline{x}) \notin [X^2_{n,\alpha/2}, X^2_{n,1-\alpha/2}]$$

où $X^2_{n,u}$ est le quantile d'ordre u pour une distribution chi-carrée à n degrés de liberté.

9.7 Test de la variance d'une population normale d'espérance inconnue

Conditions : Soit $\underline{X} = (X_1, X_2, ..., X_n)$ un échantillon aléatoire simple, issu d'une population $N(\mu, \sigma^2)$, avec μ inconnue et soit $\underline{x}$ un échantillon réalisé.

Hypothèses : Nous réalisons un test *bilatéral*. Nous désirons donc choisir entre

$$H_0 : \sigma = \sigma_0$$
$$H_1 : \sigma \neq \sigma_0$$

Statistique de test :

$$T(\underline{X}) = \frac{(n-1)\hat{S}^2_{n,c}}{\sigma_0^2}$$

Loi sous H_0 : Nous savons que

$$T(\underline{X}) =_d X^2_{n-1}$$

Règle de comportement : Il faut *rejeter l'hypothèse nulle*, avec un risque α lorsque la statistique calculée sur l'échantillon réalisé $\underline{x}$, soit :

$$T(\underline{x}) = \frac{(n-1)\hat{S}^2_{n,c}(\underline{x})}{\sigma_0^2}$$

est telle que

$$T(\underline{x}) \notin [X^2_{n-1,\alpha/2}, X^2_{n-1,1-\alpha/2}]$$

où $X^2_{n,u}$ est le quantile d'ordre u pour une distribution chi-carrée à n degrés de liberté.

9.8 Test de comparaison des variances issues de deux populations de loi normale

Conditions : Soit
- $\underline{X} = (X_1, X_2, ..., X_{n_1})$ un échantillon aléatoire simple, issu d'une population de $N(\mu_1, \sigma_1^2)$.
- $\underline{Y} = (Y_1, Y_2, ..., Y_{n_2})$ un échantillon aléatoire simple, issu d'une population de $N(\mu_2, \sigma_2^2)$.

On suppose également que ces deux échantillons sont indépendants entre eux.

Hypothèses : Nous réalisons un test *bilatéral*. Nous désirons donc choisir entre :

$$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$$
$$H_1 : \frac{\sigma_1^2}{\sigma_1^2} \neq 1$$

Statistique de test :

$$T(\underline{X}, \underline{Y}) = \frac{\hat{S}^2_{X,c}}{\hat{S}^2_{Y,c}}$$

Loi sous H_0 : On peut démontrer que

T($\underline{X}$, $\underline{Y}$) $=_d$ Fisher-Snedecor de $n_1 - 1$ et $n_2 - 1$ degrés de liberté.

Règle de comportement : Il faut *rejeter l'hypothèse nulle*, avec un risque α lorsque la statistique calculée sur les échantillons réalisés $\underline{x}$, $\underline{y}$.

$$T(\underline{x}, \underline{y}) = \frac{\hat{S}^2_{n_1,c}(\underline{x})}{\hat{S}^2_{n_2,c}(\underline{y})}$$

est telle que

$$T(\underline{x}, \underline{y}) \notin [f_{(n_1-1,n_2-1),\alpha/2}, f_{(n_1-1,n_2-1),1-\alpha/2}]$$

où $f_{(n,p),u}$ est le quantile d'ordre u pour une distribution de Fisher-Snedecor avec (n, p) degrés de liberté.

Test de comparaison des variances issues de deux populations de loi normale de variance inconnue dans R : On va utiliser
`var.test(vecteur1, vecteur2, alternative?=, conf.level?=)`.

9.9 Test d'indépendance

Conditions : Soit un échantillon de n individus (avec $n \geq 50$) sur lequels nous mesurons X et Y, ainsi nous obtenons :

— $\underline{X} = (X_1, X_2, ..., X_n)$ un vecteur d'observation de X.
— $\underline{Y} = (Y_1, Y_2, ..., Y_n)$ un vecteur d'observation de Y.

Théoriquement, X et Y sont telles qu'elles présentent respectivement n_p et n_q valeurs ou catégories possibles, notées $C_{X,1}, C_{X,2}, ..., C_{X,p}$ et $C_{Y,1}, C_{Y,2}, ..., C_{Y,q}$.
De plus, théoriquement, nous savons que :

$$\begin{aligned} Pr[X \in C_{X,i}, Y \in C_{Y,j}] &= p_{ij} \\ Pr[X \in C_{X,i}] &= p_{i\bullet} \\ Pr[Y \in C_{Y,j}] &= p_{\bullet j} \end{aligned}$$

Soit deux échantillons réalisés $\underline{x}$ et $\underline{y}$ tels que l'effectif observé pour chaque couple de catégories, soit $n_{i,j}$ *est supérieur ou égal à 5.*

Hypothèses : Nous désirons choisir entre :

— H_0 : les variables X et Y sont indépendantes : $p_{ij} = p_{i\bullet}p_{\bullet j}, \forall i \in \{1, ..., p\}, j \in \{1, ..., q\}$.
— H_1 : les variables X et Y ne sont pas indépendantes : $\exists i \in \{1, ..., p\}, j \in \{1, ..., q\} : p_{ij} \neq p_{i\bullet}p_{\bullet j}$.

Statistique de test : Sur l'échantillon, on observe l'échantillon suivant représenté sous forme d'un tableau de contingence.

X/Y	$C_{Y,1}$	...	$C_{Y,j}$	...	$C_{Y,q}$	Totaux
$C_{X,1}$	$N_{1,1}$	...	$N_{1,j}$	...	$N_{1,q}$	$N_{1,\bullet}$
⋮	⋮		⋮		⋮	⋮
$C_{X,i}$	$N_{i,1}$	...	$N_{i,j}$	...	$N_{i,q}$	$N_{i,\bullet}$
⋮	⋮		⋮		⋮	⋮
$C_{X,p}$	$N_{p,1}$	...	$N_{p,j}$	...	$N_{p,q}$	$N_{p,\bullet}$
Totaux	$N_{\bullet,1}$	...	$N_{\bullet,j}$	...	$N_{\bullet,q}$	$N_{\bullet,\bullet}$

La statistique est :

$$T(\underline{X}, \underline{Y}) = \sum_{i=1}^{p}\sum_{j=1}^{q} \frac{(N_{ij} - np_{i\bullet}p_{j\bullet})^2}{np_{i\bullet}p_{\bullet j}}$$

Loi sous H_0 : Lorsque l'hypothèse H_0 est vrai, nous avons :

$$N_{ij} = np_{i\bullet}p_{\bullet j}$$

Ainsi, on peut démontrer que :

$$T(\underline{X}, \underline{Y}) =_d X^2_{(p-1)(q-1)}$$

Règle de comportement : Nous avons les observations suivantes

X / Y	$C_{Y,1}$	...	$C_{Y,j}$	...	$C_{Y,q}$	Totaux
$C_{X,1}$	$n_{1,1}$	...	$n_{1,j}$	...	$n_{1,q}$	$n_{1,\bullet}$
⋮	⋮		⋮		⋮	⋮
$C_{X,i}$	$n_{i,1}$	...	$n_{i,j}$	...	$n_{i,q}$	$n_{i,\bullet}$
⋮	⋮		⋮		⋮	⋮
$C_{X,p}$	$n_{p,1}$	...	$n_{p,j}$	...	$n_{p,q}$	$n_{p,\bullet}$
Totaux	$n_{\bullet,1}$	...	$n_{\bullet,j}$	...	$n_{\bullet,q}$	$n_{\bullet,\bullet}$

où on comptabilise les observations pour chaque modalités (X, Y) possible. Il faut *rejeter l'hypothèse nulle*, avec un risque α, lorsque :

$$T(\underline{x}, \underline{y}) > X^2_{(p-1)(q-1),1-\alpha}$$

où $X^2_{n,u}$ est le quantile d'ordre u pour une distribution de chi-carrée avec n degrés de liberté.

Test d'indépendance dans R : On va utiliser `chisq.test(tableau)`.

Exemple avec un exercice de l'examen de Janvier 2021 :
Peut-on affirmer avec 8% de chance de se tromper que fumer à un impact sur la gravite de la maladie.

Choix des hypothèses :
H_0 : Fumer n'a pas d'impact sur la maladie.
H_1 : Fumer a un impact sur la maladie

Calculs avec R :

```
# Lecture du fichier
BD<-read.csv("path-to-file", sep=" ")
# Création d'un tableau contingence
fumer_vs_gravite = table(BD$tabac,BD$gravite)
# Test d'indépendance
chisq.test(fumer_vs_gravite)
# Cette commande va retourner une p-value entre 0 et 1. Si cette p-value est plus petite que notre chance de se tromper, alors on peut rejeter H0. Sinon, on ne peut pas la rejeter.
# Par exemple, si p-value=0.53 alors on ne peut pas rejeter H0 car 0.53 > 0.08.
```

9.10 Test d'ajustement

Conditions : Soit un échantillon de n individus sur lesquels nous mesurons X, ainsi nous observons $\underline{X} = (X_1, X_2, ..., X_n)$ un vecteur d'observation de X et soit un échantillon réalisé $\underline{x}$.
Théoriquement, X est telle qu'elle présente p valeurs ou catégories possibles, notées $C_{X,1}, C_{X,2}, ..., C_{X,p}$.

Hypothèses : Nous désirons donc choisir entre :

— H_0 : la variable X suit une distribution F.
— H_1 : la variable X ne suit pas une distribution F.

Statistique de test : Sous l'hypothèse nulle, la variable aléatoire X est la loi F, soit :

$$Pr[X \in C_{X,i}] = p_i$$

pour $1 \leq i \leq p$.
Sur l'échantillon, on observe que N_i individus présentent le caractère i, pour $1 \leq i \leq p$. La statistique est alors :

$$T(\underline{X}) = \sum_{i=1}^{p} \frac{(N_i - np_i)^2}{np_i}$$

Loi sous H_0 : Lorsque l'hypothèse H_0 est vrai, nous avons :

$$N_i = np_i$$

Ainsi, on peut démontrer que :

$$T(\underline{X}) =_d X^2_{p-1}$$

Règle de comportement : Soit n_i, $1 \leq i \leq p$, le nombre d'individus observés dans la catégorie $C_{X,i}$.
Il faut *rejeter l'hypothèse nulle*, avec un risque α lorsque :

$$T(\underline{x}) > X^2_{p-1,1-\alpha}$$

où $X_{n,u}$ est le quantile d'ordre u pour une distribution de chi-carrée avec n degrés de liberté.

Condition d'utilisation de cette règle :
— $n \geq 30$
— $np_i \geq 1 \ \forall i \in \{1, ..., p\}$
— $np_i \geq 5$ dans 80% des cas au moins
Si une des trois conditions n'est pas satisfaite, nous devons regrouper des catégories de X.

Test d'ajustement dans R : On va utiliser
`chisq.test(observation_nb_vecteur, p=theorique_proba_vecteur)`.

Exemple avec un exercice de l'examen de Janvier 2021 :
Peut-on affimer avec 9% de chance de se tromper que la représentation homme/femme est issue d'une Bernouilli avec p = 0.48.

Choix des hypothèses :
H_0 : la variable sexe ne suit pas une loi de Bernouilli de paramètre $p = 0.48$.
H_1 : la variable sexe suit une loi de Bernouilli de paramètre $p = 0.48$.

Calculs avec R :

```
# Lecture du fichier
BD<-read.csv("path-to-file", sep=" ")

# Création d'un vecteur contenant le nombre d'hommes et de femmes observés
observation = as.vector(table(BD$sexe))

# Création d'un vecteur contenant la probabilité théorique
# On veut savoir si la représentation homme/femme est issue d'une Bernouilli avec p = 0.48, donc 48% homme et 52% femme
# On met 0.52 à gauche car dans le vecteur observation, les femmes sont à gauche.
theorique = c(0.52, 0.48)

# Test d'ajustement
chisq.test(observation, p=theorique)

# Cette commande va retourner une p-value entre 0 et 1. Si cette p-value est plus petite que notre chance de se tromper, alors on peut rejeter H0. Sinon, on ne peut pas la rejeter.
# Par exemple, si p-value=0.003 alors on peut rejeter H0 car 0.003 < 0.09.
```