Tests d'hypothèses (généralités - test de conformité)

Tewfik Mahdjoub

Université Abou Bekr Belkaïd, Tlemcen

Novembre 2024

Plan du cours

Chapitre 1 : Analyse d'une série statistique

Groupes	Fe filet	Cd filet	Ni filet	Co filet	Pb filet	Fe branchies	Cd branchies	Ni branchies	Co branchies	Pbbranchies	Fe estomac	Cd estomac	Ni estomac	Co estomac	Pbestomac
1	1,144	0,025	0,615	0,022	0,012	2,147	0,049	0,379	0,015	0,280	2,928	0,039	1,160	0,029	0,011
2	1,532	0,092	1,705	0,015	0,517	6,109	0,090	0,742	0,049	1,069	4,609	0,054	0,375	0,023	0,219
3	1,182	0,033	1,588	0,019	0,149	3,504	0,069	2,085	0,037	0,032	1,483	0,000	2,360	0,021	0,031
4	2,155	0,012	2,626	0,015	0,306	6,533	0,073	1,866	0,032	1,113	5,708	0,048	2,278	0,015	0,554
5	2,141	0,081	2,808	0,030	1,905	2,266	0,074	3,025	0,024	1,007	1,560	0,036	3,358	0,023	0,542
6	1,796	0,098	3,221	0,031	2,008	2,556	0,162	3,007	0,042	1,753	3,116	0,097	3,728	0,029	1,176
7	1,739	0,081	2,249	0,015	1,320	5,427	0,119	4,363	0,041	2,420	3,860	0,110	4,355	0,013	1,134
8	0,974	0,098	3,550	0,075	2,290	3,395	0,134	3,124	0,114	2,777	3,435	0,093	5,380	0,094	0,919
9	1,293	0,131	4,706	0,0[illegible]			[illegible]0	3,857	0,096	4,770	1,941	0,101	3,923	0,068	2,284
10	2,436	0,178	5,086	0,1[illegible]			[illegible]4	4,716	0,104	7,087	3,001	0,182	5,632	0,100	4,371
11	1,264	0,130	3,918	0,1[illegible]			0,156	4,297	0,128	8,336	1,742	0,122	3,385	0,119	0,502
12	1,158	0,171	4,669	0,074	21,270	3,515	0,174	4,696	0,108	6,842	1,719	0,145	4,832	0,071	4,576
13	1,313	0,258	6,695	0,061	0,004	3,270	0,261	6,706	0,088	0,152	1,625	0,214	6,656	0,065	0,000
14	1,237	0,204	7,523	0,072	0,319	3,810	0,276	6,852	0,074	0,650	1,090	0,198	7,844	0,067	0,152
15	1,260	0,193	7,467	0,120	1,196	2,726	0,223	7,978	0,110	1,777	2,557	0,200	8,247	0,105	1,147
16	1,520	0,236	8,631	0,113	0,618	3,133	0,299	8,485	0,173	1,989	2,859	0,257	8,847	0,125	0,948
17	1,883	0,275	8,678	0,091	1,181	4,656	0,302	9,454	0,136	1,440	4,339	0,286	8,187	0,094	0,987
18	1,813	0,258	9,532	0,108	2,088	3,564	0,320	9,622	0,098	3,144	1,144	0,288	9,409	0,081	1,863
19	1,578	0,263	10,880	0,188	2,498	2,653	0,285	10,230	0,168	2,884	1,703	0,252	9,495	0,171	2,033
20	1,596	0,317	11,720	0,160	2,276	3,635	0,301	11,420	0,207	4,704	1,871	0,271	11,620	0,145	1,610
21	1,832	0,304	11,210	0,091	1,687	3,340	0,362	12,600	0,145	3,185	1,774	0,321	10,970	0,105	1,479
22	1,907	0,348	12,990	0,050	3,873	4,786	0,374	12,850	0,079	11,030	2,018	0,321	12,290	0,041	3,166

Echantillon $\mathcal{E}$

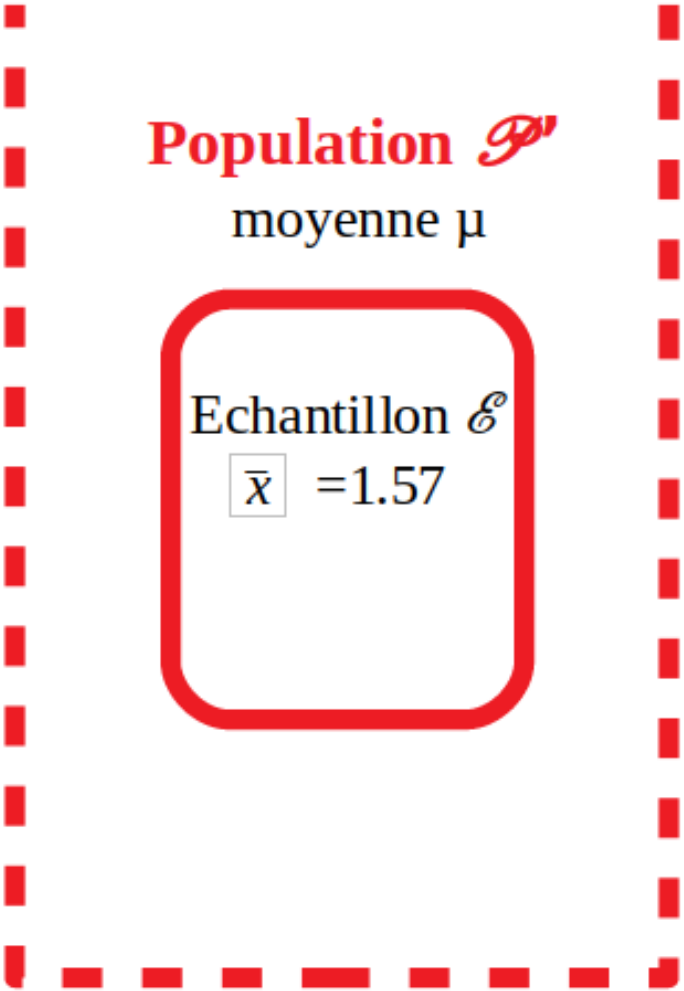
Tableau 1 :Concentration des métaux lourds au niveau du filet, des branchies et de l'estomac (mg/kg de poids sec)

$\bar{x} = 1.57$ mg/kg de poids sec $\quad \sigma_x = 0.33$ mg/kg de poids sec

Bouali, M., and Kedrouci, A., 2006. Bioaccumulation de cinq métaux lourds (Fe, Ni, Cd, Pb, Co) au niveau des branchies, de l'estomac et du filet chez la sardine (*Sardina pilchardus*) dans la baie de Béni-Saf. Application de la régression multiple. Mémoire d'ingénieur d'état en écologie animale.

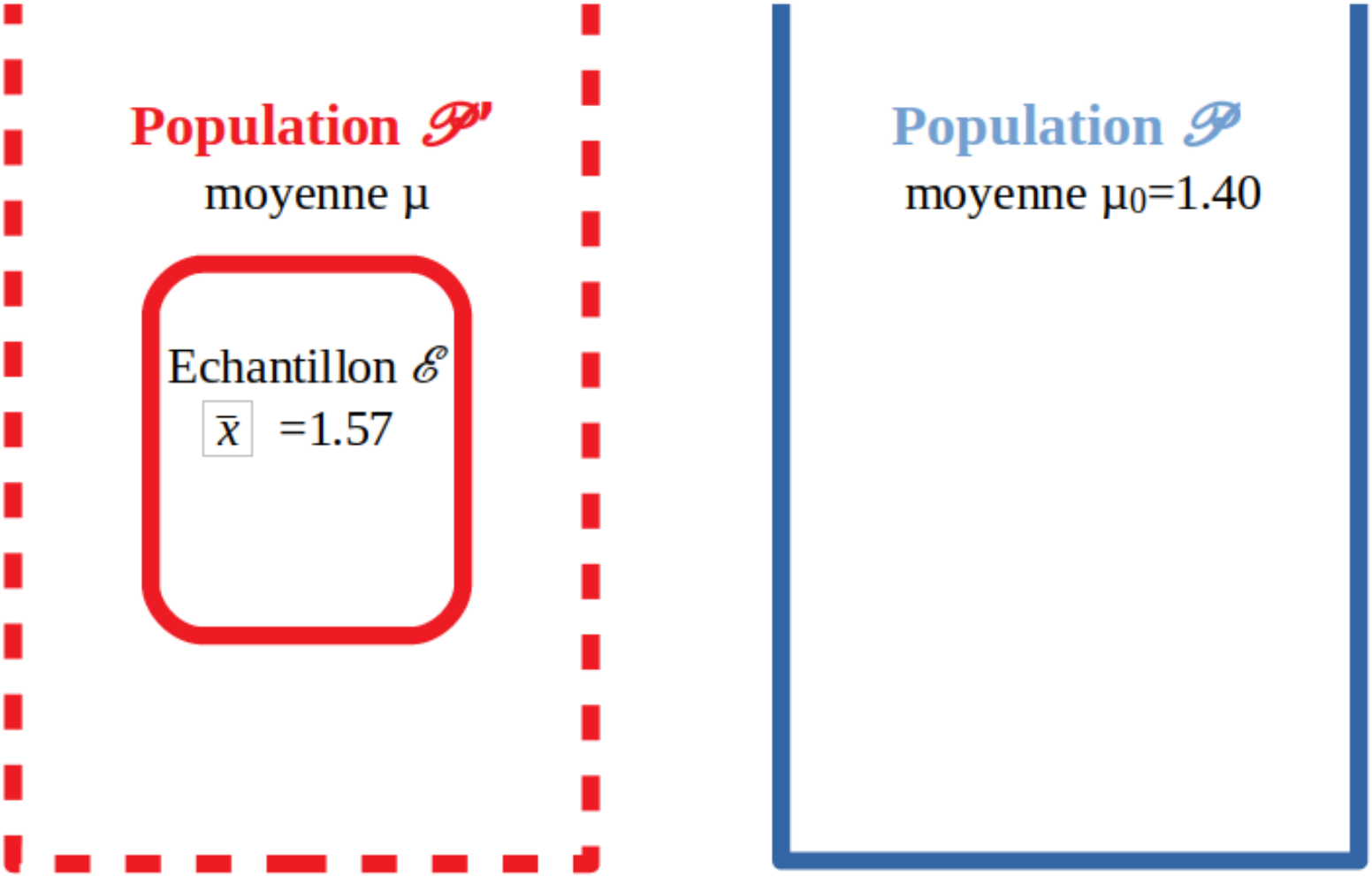
Echantillon $\mathscr{E}$

$\bar{x}$ =1.57



$\mathscr{P}'$: population d'où est extrait l'échantillon $\mathscr{E}$

μ : moyenne inconnue de la population $\mathscr{P}'$



$\mathscr{P}$: population de référence

μ_0 : moyenne imposée par le législateur

Question :
le taux de fer dans le filet des sardines de la zone de Béni-Saf (population $\mathscr{P}'$) est-il acceptable ?

autrement dit :

la population $\mathscr{P}'$ a-t-elle la même moyenne que la population $\mathscr{P}$?

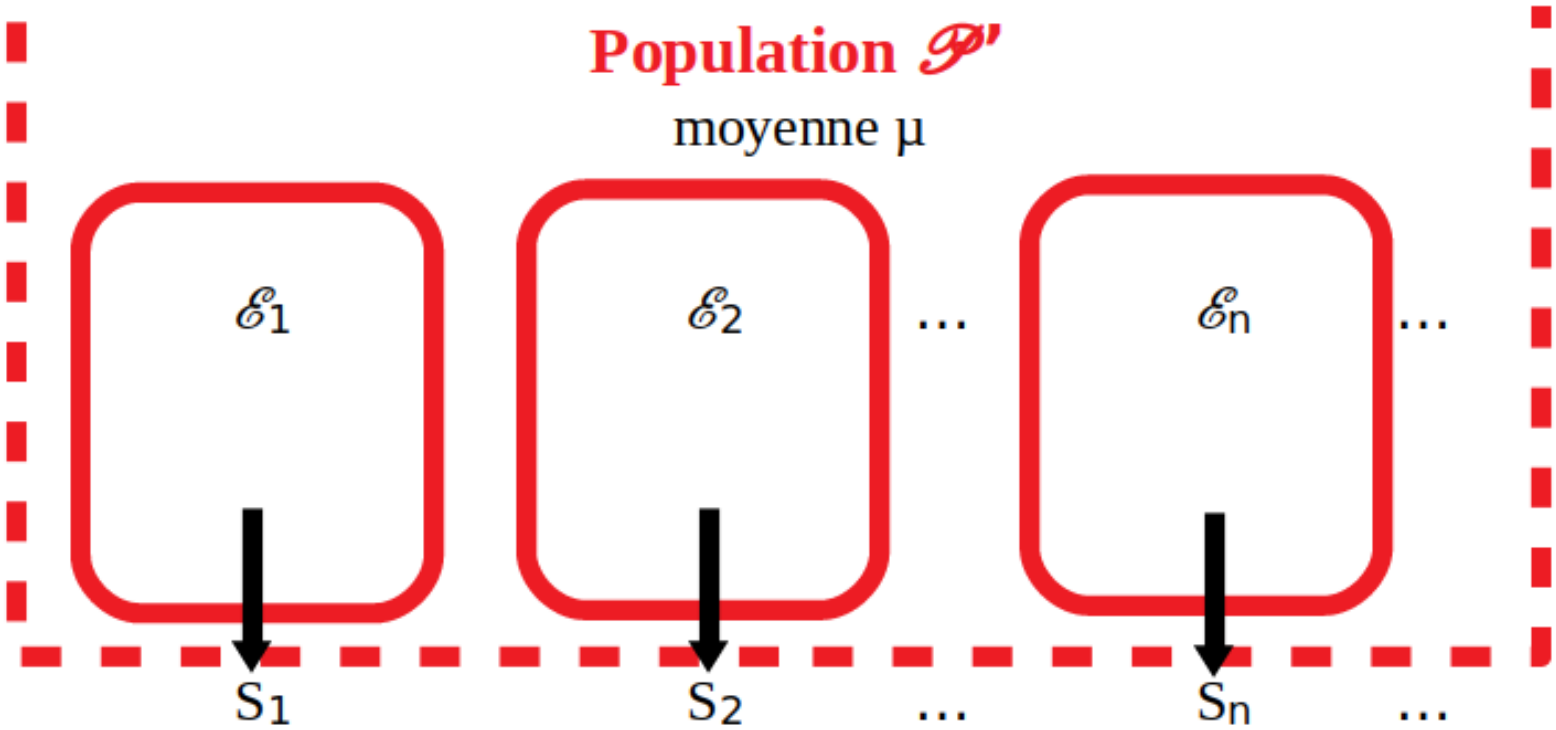
$$\mu = \mu_0 \, ?$$

Pour répondre à cette question, on calcule une statistique :

$$S = S(\mathscr{E})$$

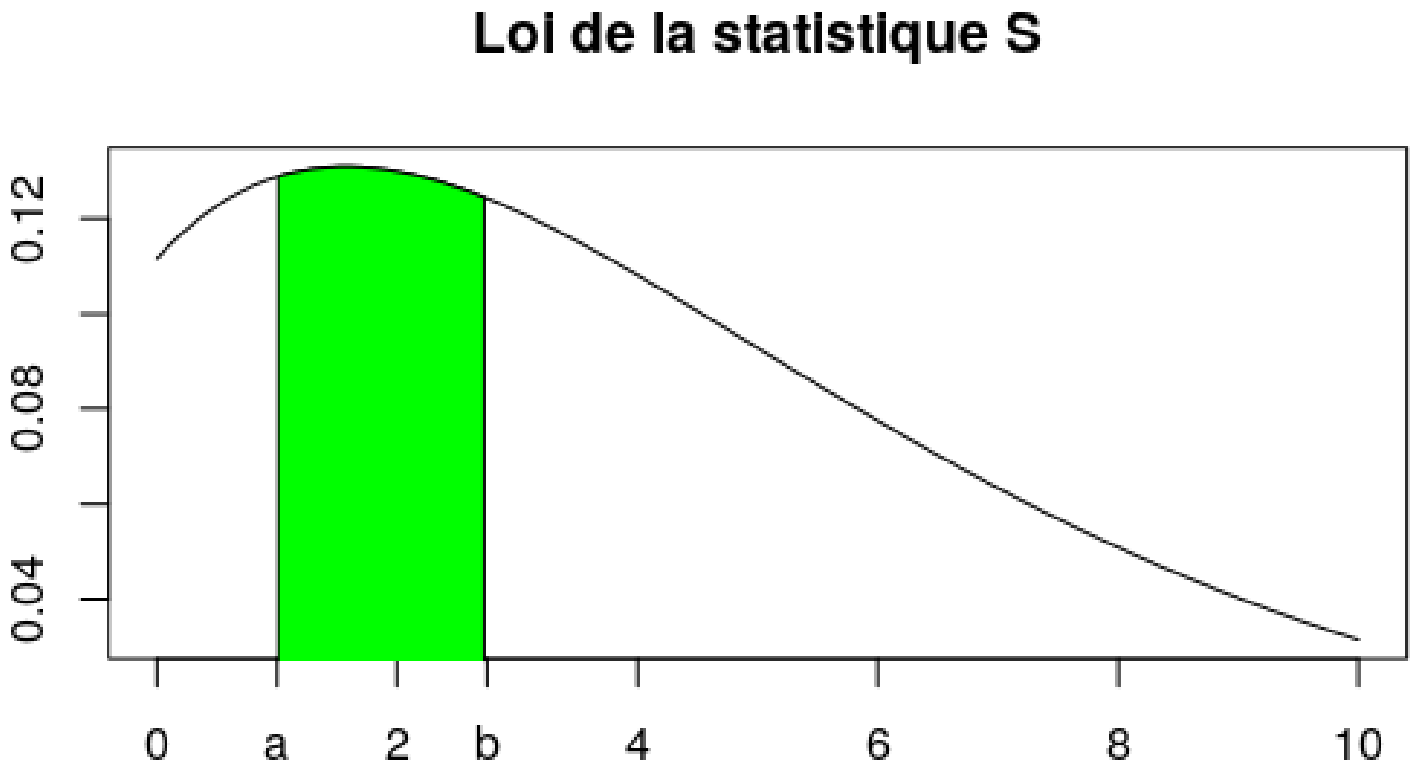
De façon générale, pour le calcul de S :

- soit il y a des différences $\Rightarrow$ on compare S à 0
- soit il y a un rapport $\Rightarrow$ on compare S à 1



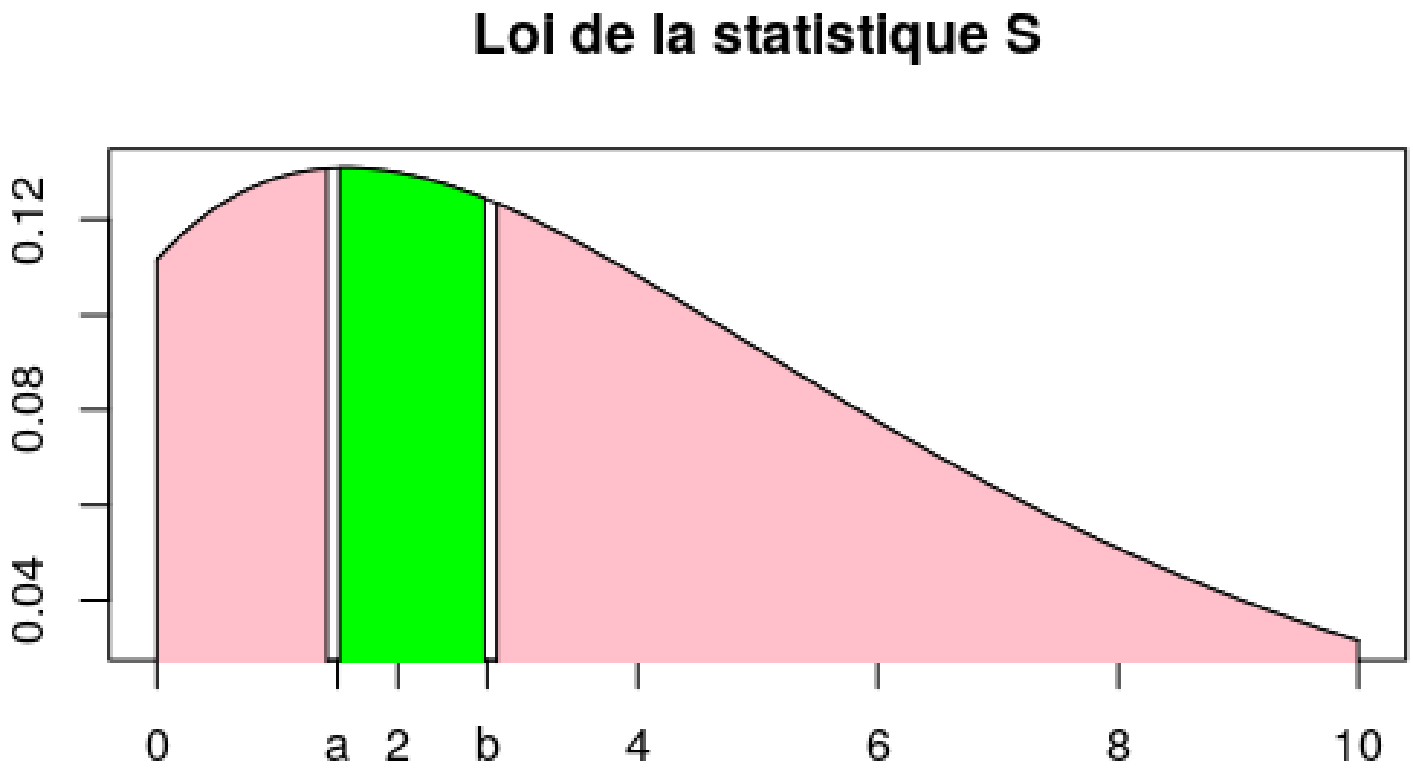
$\{S_1, S_2, ..., S_n, ...\}$ sont donc des valeurs (d'une variable aléatoire) qui suivent une certaine loi, de fonction densité de probabilité f.

Décision : les valeurs de $S \in [a, b] \Rightarrow \mu = \mu_0$



Rappel : $P(a \leq S \leq b) = \int_a^b f(x)dx$=aire hachurée en vert

Risque :$\exists\ S^* \notin [a, b]$ telle que la statistique S^* correspondant à l'échantillon $\mathscr{E}^*$ extrait de la population $\mathscr{P}'$.



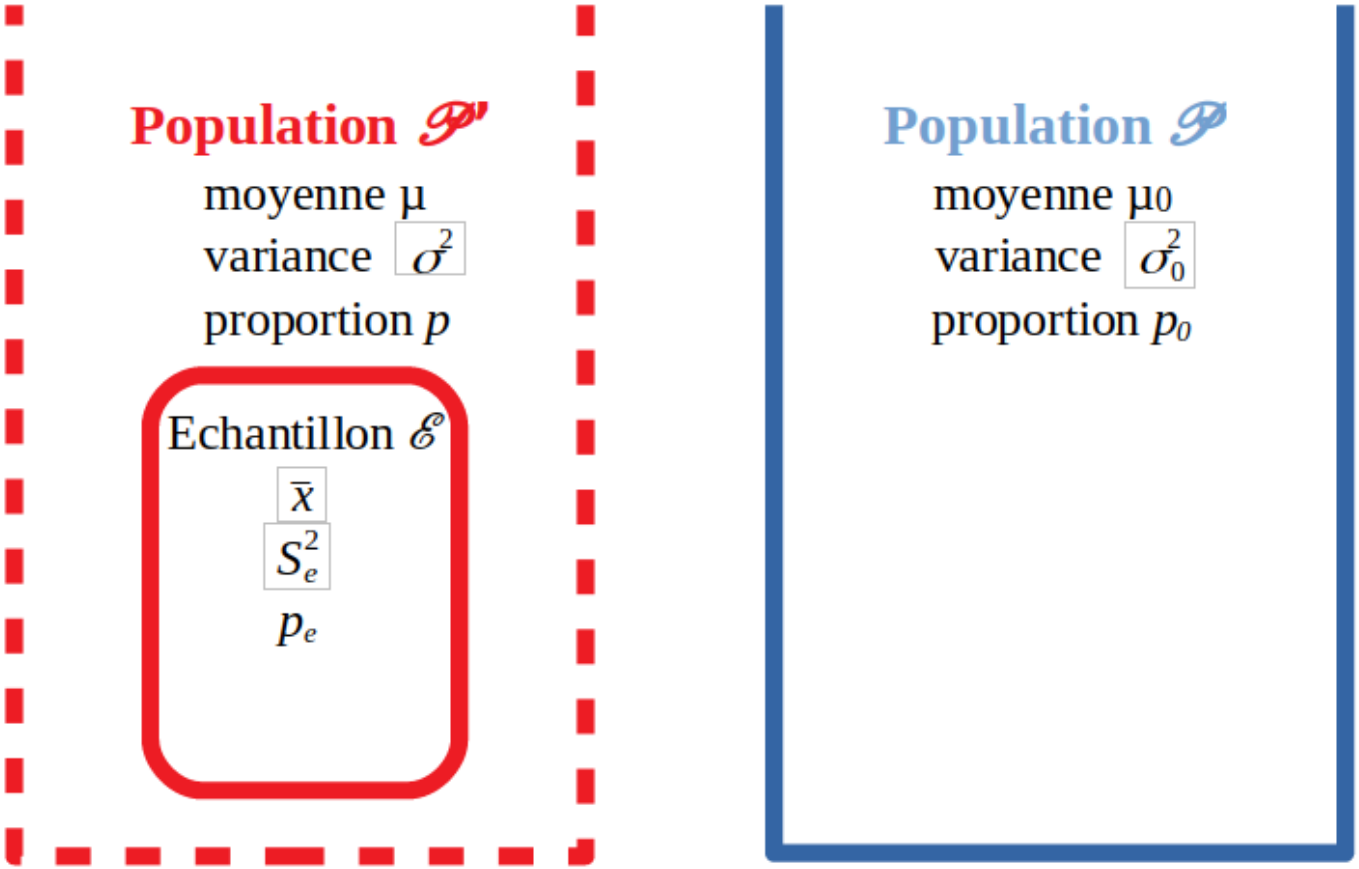
Définition 1 : le risque de rejeter des échantillons alors qu'ils sont extraits de la population $\mathscr{P}'$ est appelé un risque d'erreur de première espèce. Il est noté α (alpha). Il est mesuré par l'aire hachurée en rose

Définition 2 :
Un test statistique est une procédure de décision sur l'acceptation ou le rejet d'une hypothèse statistique, appelée hypothèse nulle H_0.

Dans le cas du problème précédent, $H_0 : \mu = \mu_0$.

Il existe 4 types de tests :

- test de conformité
- test d'homogénéité
- test d'ajustement (d'adéquation)
- test d'indépendance (d'association ou de corrélation)

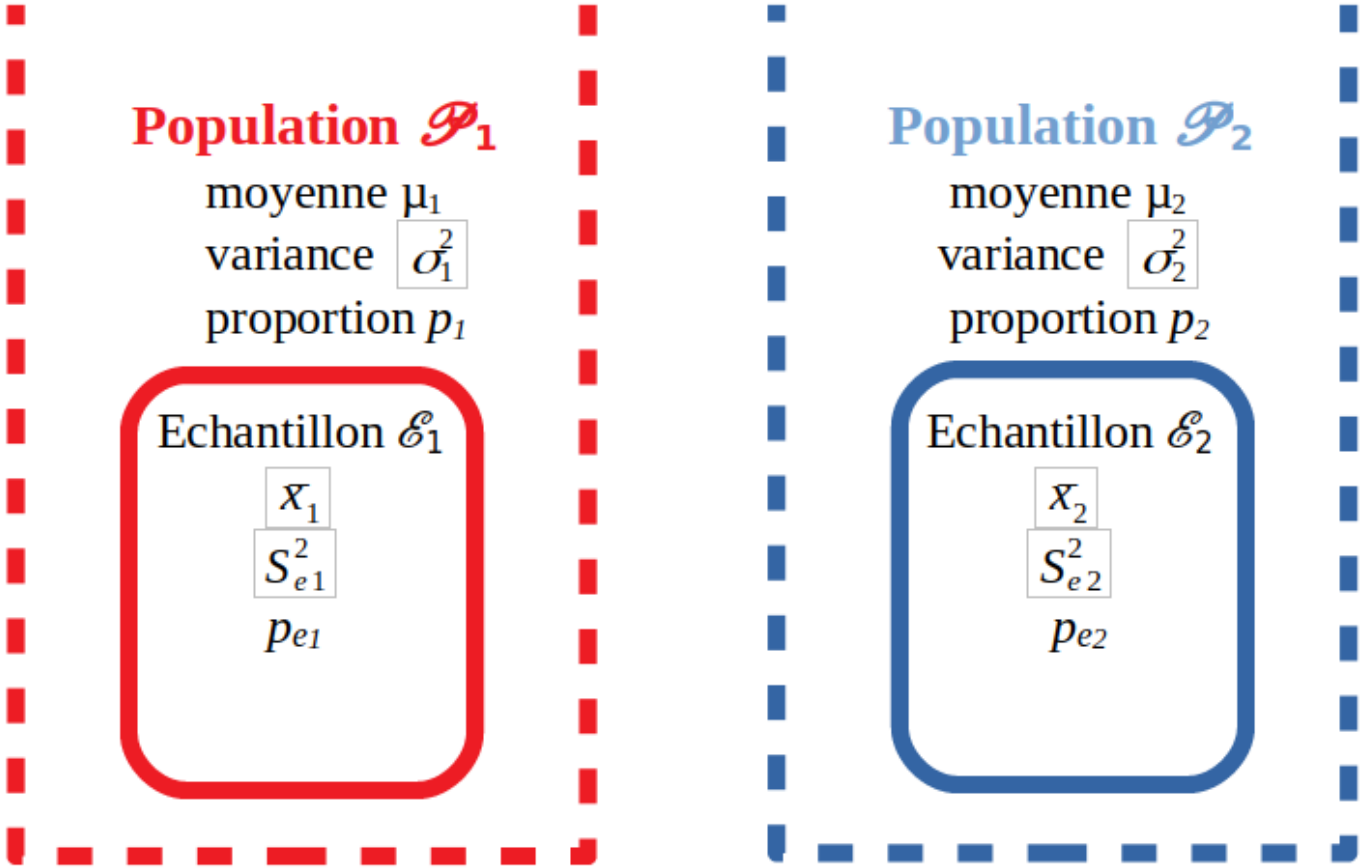

Population $\mathscr{P}'$
moyenne μ
variance σ^2
proportion p
Echantillon $\mathscr{E}$
$\bar{x}$
S_e^2
p_e
Population $\mathscr{P}$
moyenne μ_0
variance σ_0^2
proportion p_0

Variable statistique quantitative :

$$H_0 : \mu = \mu_0 \quad ou \quad \sigma^2 = \sigma_0^2$$

Variable statistique qualitative :

$$H_0 : p = p_0$$

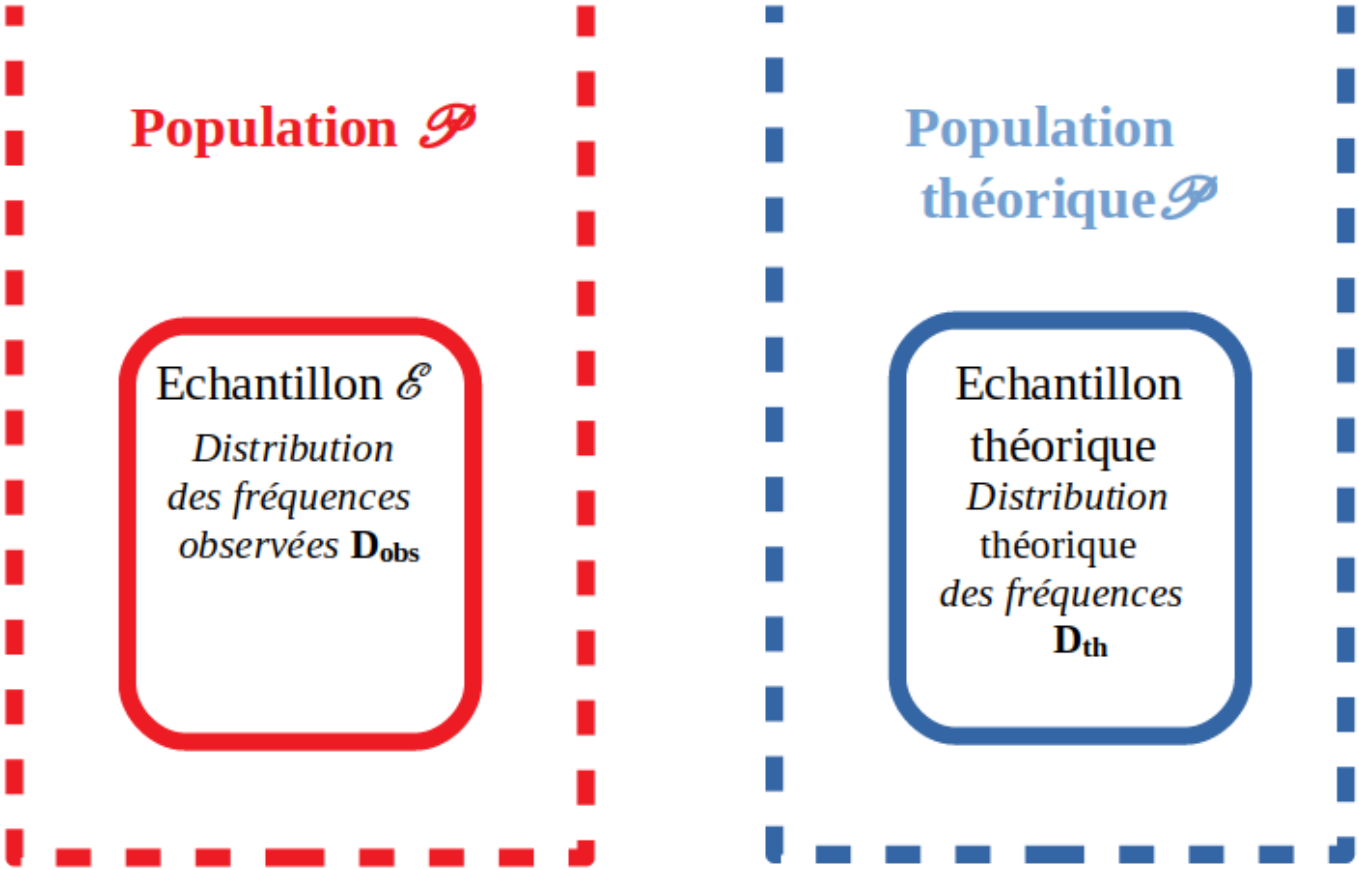

Population $\mathscr{P}_1$
moyenne μ_1
variance σ_1^2
proportion p_1
Echantillon $\mathscr{E}_1$
x_1
S_{e1}^2
p_{e1}
Population $\mathscr{P}_2$
moyenne μ_2
variance σ_2^2
proportion p_2
Echantillon $\mathscr{E}_2$
x_2
S_{e2}^2
p_{e2}

Variable statistique quantitative :

$$H_0 : \mu_1 = \mu_2 \quad ou \quad \sigma_1^2 = \sigma_2^2$$

Variable statistique qualitative :

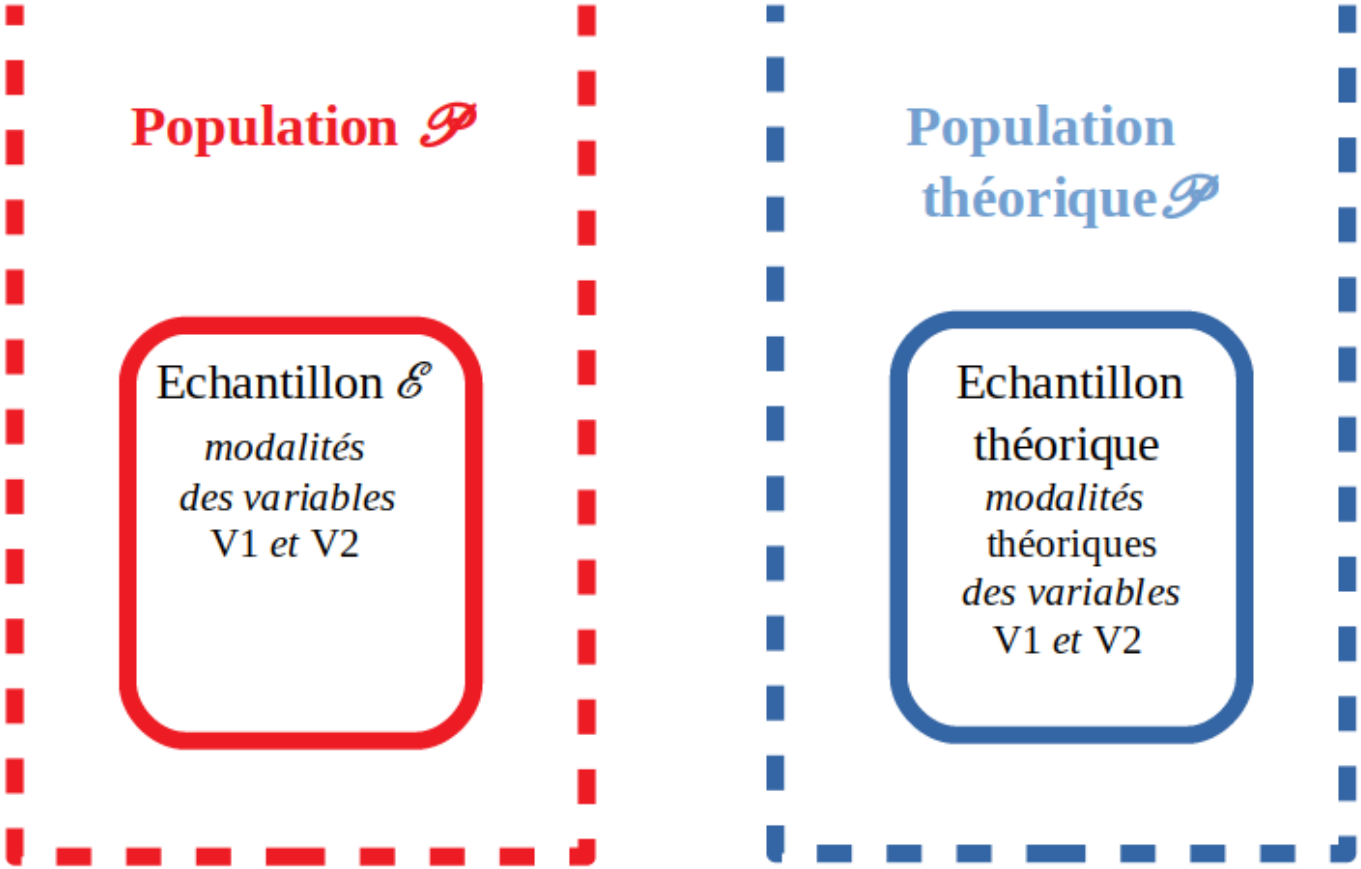
$$H_0 : p_1 = p_2$$


Population $\mathscr{P}$
Echantillon $\mathscr{E}$
Distribution des fréquences observées $\mathbf{D_{obs}}$
Population théorique $\mathscr{P}$
Echantillon théorique
Distribution théorique *des fréquences*
$\mathbf{D_{th}}$

Test d'ajustement

Variable statistique quantitative :

$$H_0 : D_{obs} = D_{th}$$


Population $\mathscr{P}$
Echantillon $\mathscr{E}$
modalités
des variables
V1 *et* V2
Population
théorique $\mathscr{P}$
Echantillon
théorique
modalités
théoriques
des variables
V1 *et* V2

Variable statistique qualitative :

$$H_0 : V_1 \text{ et } V_2 \text{ indépendants}$$

Hypothèse alternative H_1

On appelle hypothèse alternative H_1 toute hypothèse différente de l'hypothèse nulle.

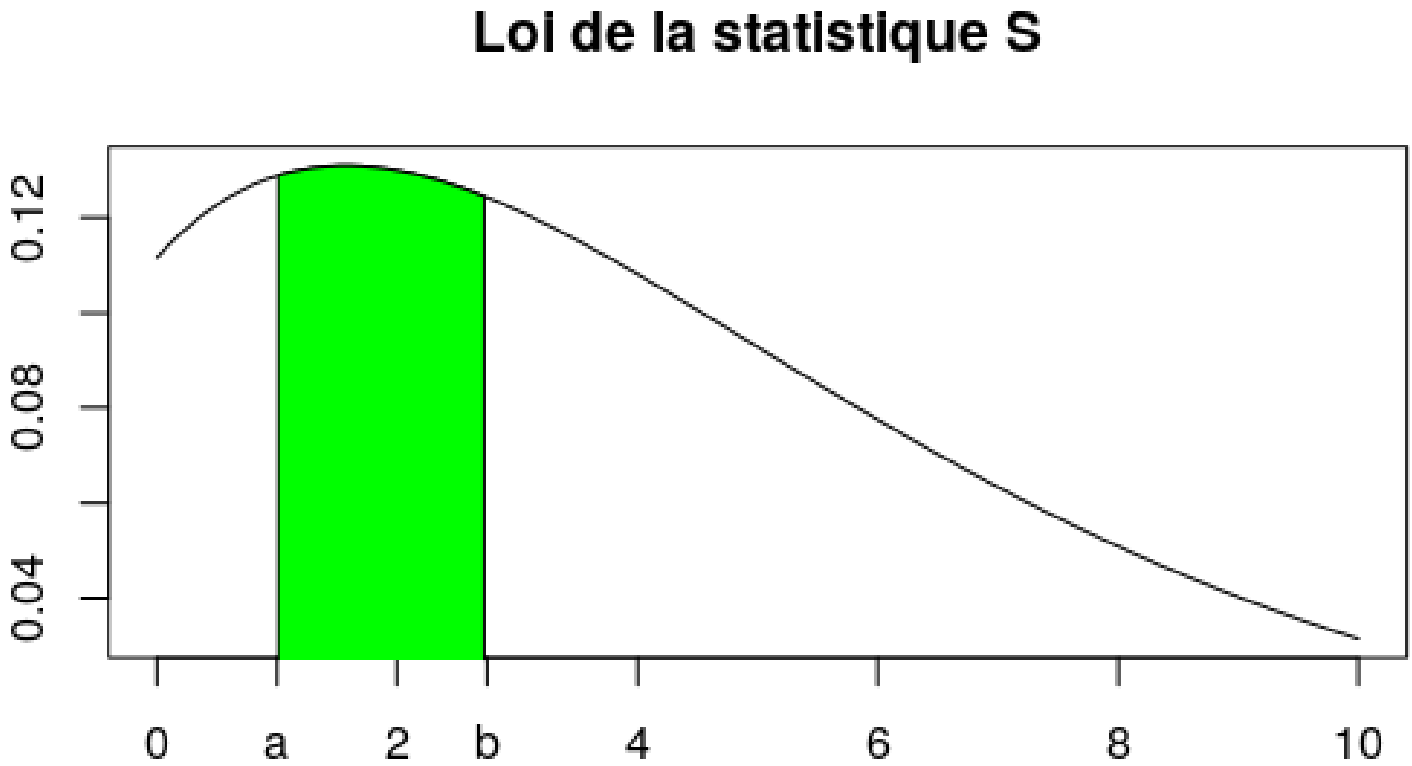
Exemple :
Si $H_0 : \mu = \mu_0$ alors $H_1 : \mu \neq \mu_0$

Remarque importante :
$H_1 : \mu < \mu_0$ ou $H_1 : \mu > \mu_0$ peuvent être aussi des hypothèses alternatives à l'hypothèse nulle H_0

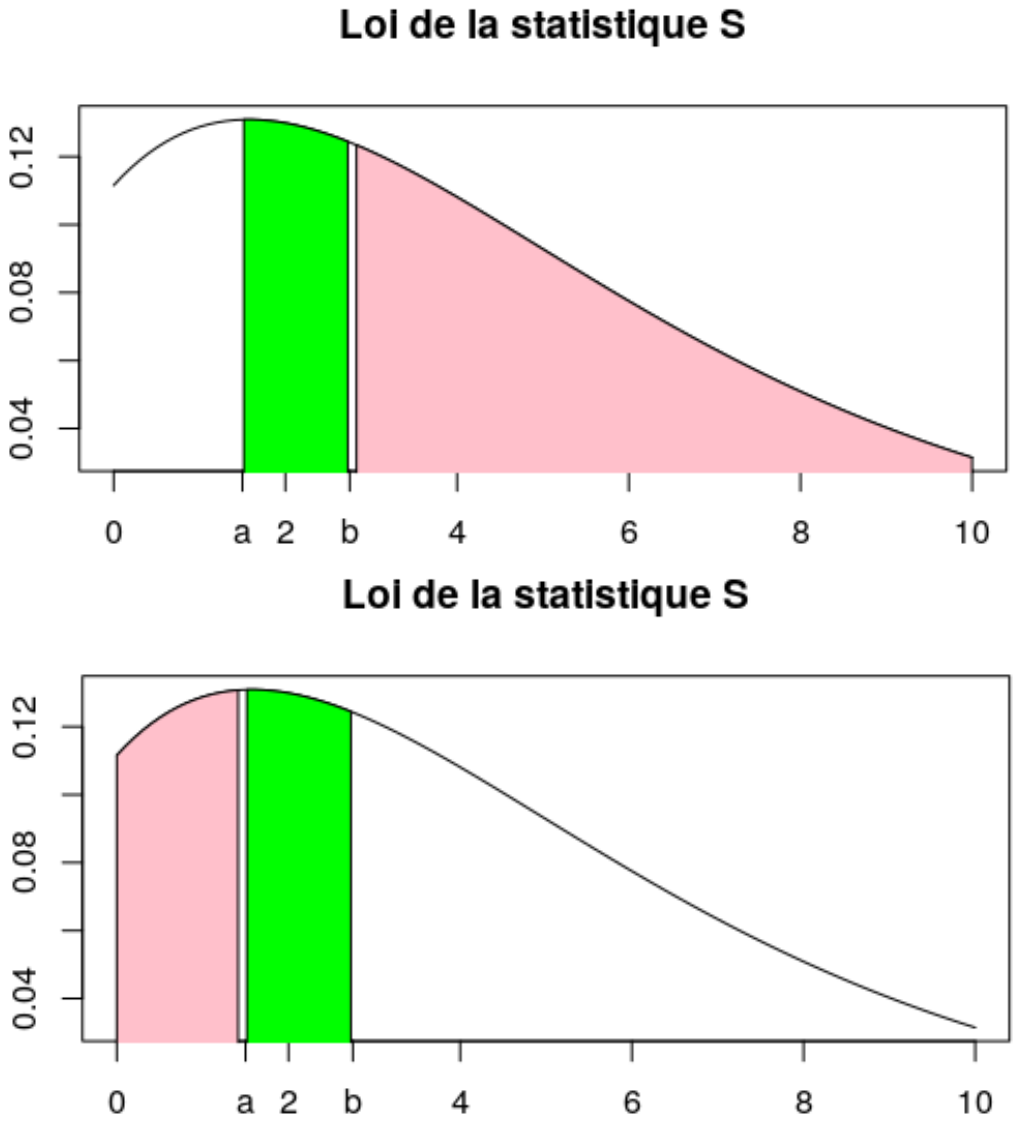
Question : comment choisir la bonne hypothèse H_1 ?
Réponse : à partir des informations que nous avons à apriori sur l'échantillon (ou les échantillons).

Décision : les valeurs de $S \in [a, b] \Rightarrow \mu = \mu_0$

Dans le cas où on n'a pas d'information sur H_1 ($\mu < \mu_0$ *ou* $\mu > \mu_0$), on **partage** le risque α de se tromper à gauche et à droite de $[a, b]$: c'est un test bilatéral.



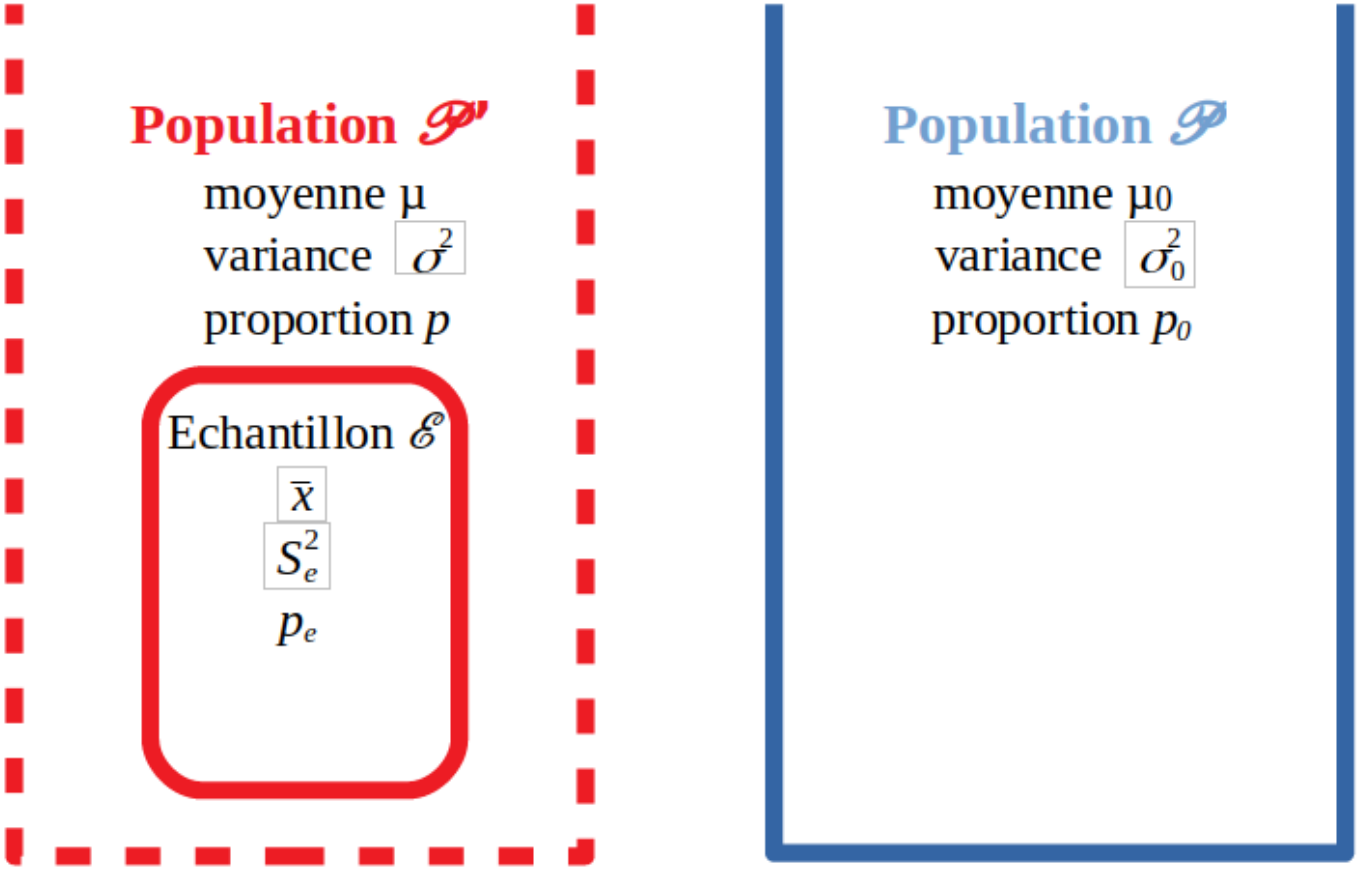
Dans le cas où on a des informations à priori sur H_1, le risque total α est mis du côté du sens de l'inégalité dans H_1 :



Un test statistique se compose de 5 étapes :

- poser H_0 et fixer le risque α
- poser H_1 et voir si le test est unilatéral ou bilatéral
- calculer la statistique S qui sera définie en fonction de la nature des tests
- comparer la probabilité de rejet de H_0 (p-value) au risque α
- conclure

Rappel :



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

Deux cas sont considérés : σ^2 connue et σ^2 inconnue

Conditions d'application du test :

- la variable statistique est quantitative
- l'échantillon est aléatoire simple
- la variance de la population est connue
- la population est normalement distribuée ou la taille de l'échantillon est supérieure à 30

Remarque : condition sur la distribution de la population⇒**test paramétrique**.

La statistique du test est :

$$S = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \tag{1}$$

S suit une loi normale dnorm(0,1)

Conditions d'application du test :

- la variable statistique est quantitative
- l'échantillon est aléatoire simple
- la variance de la population est inconnue
- la population est normalement distribuée ou la taille de l'échantillon est supérieure à 30

La statistique du test est :

$$S = \frac{\bar{x} - \mu_0}{\frac{s_e}{\sqrt{n}}} \tag{2}$$

S suit une loi de Student à n-1 degré de liberté (ddl = n-1).
On a la relation importante :

$$\sigma^2 = \frac{n}{n-1} s_e^2 \tag{3}$$

- la variable statistique est quantitative
- l'échantillon est aléatoire simple
- la population a une distribution normale

La statistique du test est :

$$S = \frac{(n-1)s_e^2}{\sigma_0^2} \tag{4}$$

S suit une loi chisq(n-1) (Khi deux à (n-1) ddl).

Exercice : Dans le tableau N° 1, intitulé "Concentration des métaux lourds au niveau du filet, des branchies et de l'estomac", on considère la 1ère colonne représentant le dosage du plomb au niveau du filet.

1. Peut-on dire, au seuil $\alpha = 5\%$, que le taux de fer moyen est égal à la norme du législateur valant 1.40 mg/kg de poids sec.
2. Utiliser le niveau de significativité de 0.05 pour tester l'affirmation que le taux de fer moyen est supérieur à la norme du législateur.
3. Tester, au seuil $\alpha = 5\%$, que la variance de la population est égale à 0.10.

>ferfilet=c(1.144, 1.532, 1.182, 2.155, 2.141, 1.796, 1.739, 0.974, 1.293, 2.436, 1.264, 1.158, 1.313, 1.237, 1.260, 1.520, 1.883, 1.813, 1.578, 1.596, 1.832, 1.907)

> t.test(ferfilet, mu=1.4, alternative="two.sided", conf.level = 0.95)

One Sample t-test	nature du test
data : ferfilet	fichier de données
t=2.1795, df=21, p-value=0.04082	t :valeur de la statistique, df :ddl, p-value :probabilité de rejet
alternative hypothesis : true mean is not equal to 1.4	hypothèse alternative
95 percent confidence interval :	intervalle de confiance à 95%
1.408234 1.751130	
sample estimates :	estimations sur l'échantillon
mean of x 1.579682	moyenne de l'échantillon

$$\textbf{p-value} = \textbf{0.04082} < \alpha = \textbf{0.05} \Rightarrow \textbf{rejet de } H_0$$

> t.test(ferfilet, mu=1.4, alternative="greater", conf.level = 0.95)

One Sample t-test	nature du test
data : ferfilet	fichier de données
t=2.1795, df=21, p-value=0.02041	t :valeur de la statistique, df :ddl, p-value :probabilité de rejet
alternative hypothesis : true mean is greater than 1.4	hypothèse alternative
95 percent confidence interval :	intervalle de confiance à 95%
1.43782 Inf	
sample estimates :	estimations sur l'échantillon
mean of x 1.579682	moyenne de l'échantillon

p-value = 0.02041 $<$ α **= 0.05** $\Rightarrow$ **rejet de** H_0

> t.test(ferfilet, mu=1.4, alternative="less", conf.level = 0.95)

One Sample t-test	nature du test
data : ferfilet	fichier de données
t=2.1795, df=21, p-value=0.9796	t :valeur de la statistique, df :ddl, p-value :probabilité de rejet
alternative hypothesis : true mean is less than 1.4	hypothèse alternative
95 percent confidence interval :	intervalle de confiance à 95%
-Inf 1.721544	
sample estimates :	estimations sur l'échantillon
mean of x 1.579682	moyenne de l'échantillon

p-value = 0.9796 $>$ α **= 0.05** $\Rightarrow$ **acceptation de** H_0

> t.test(ferfilet, mu=1.4, alternative="two.sided", conf.level = 0.99)

One Sample t-test	nature du test
data : ferfilet	fichier de données
t=2.1795, df=21, p-value=0.04082	t :valeur de la statistique, df :ddl, p-value :probabilité de rejet
alternative hypothesis : true mean is not equal to 1.4	hypothèse alternative
99 percent confidence interval :	intervalle de confiance à 99%
1.346258 1.813106	
sample estimates :	estimations sur l'échantillon
mean of x 1.579682	moyenne de l'échantillon

p-value = 0.04082 $>$ α = 0.01 $\Rightarrow$ acceptation de H_0

Charger le package **EnvStats** pour la commande varTest

> varTest(ferfilet,alternative = "two.sided",conf.level = .9, sigma.squared = 0.1)

Results of Hypothesis Test	
Null Hypothesis :	variance = 0.1
Alternative Hypothesis :	True variance is not equal to 0.1
Test Name :	Chi-Squared Test on Variance
Estimated Parameter(s) :	variance = 0.1495283
Data :	ferfilet
Test Statistic :	Chi-Squared = 31.40095
Test Statistic Parameter :	df = 21
P-value :	0.1344783
90% Confidence Interval :	LCL = 0.09611386 UCL = 0.27090088

p-value = 0.1344783 $>$ α **= 0.1** $\Rightarrow$ **acceptation de** H_0

> varTest(ferfilet,alternative = "greater", conf.level = .9, sigma.squared = 0.1)

Results of Hypothesis Test	
Null Hypothesis :	variance = 0.1
Alternative Hypothesis :	True variance is greater than 0.1
Test Name :	Chi-Squared Test on Variance
Estimated Parameter(s) :	variance = 0.1495283
Data :	ferfilet
Test Statistic :	Chi-Squared = 31.40095
Test Statistic Parameter :	df = 21
P-value :	0.06723915
90% Confidence Interval :	LCL = 0.1060302 UCL = Inf

$$\textbf{p-value} = \mathbf{0.06723915} < \alpha = \mathbf{0.1} \Rightarrow \textbf{rejet de } H_0$$

- la variable statistique est qualitative
- schéma de Bernoulli (un nombre n fixé d'essais indépendants et chaque essai n'a que 2 résultats possibles : succès (probabilité p) et échec (probabilité q)
- $np > 5$ et $nq > 5$

La statistique du test est :

$$S = \frac{p_e - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \tag{5}$$

S suit une loi normale dnorm(0,1)

Exercice : un directeur de journal affirme que plus de 80% des foyers lisent au moins un quotidien. Un sondage effectué auprès de 1000 foyers indique que 840 lisent au moins un quotidien.
Est-ce que l'affirmation du directeur est supportée par les résultats du sondage, aux seuils $\alpha = 1\%,\ 5\%$ *et* 10% ?
Solution :
> binom.test(840,1000,0.8,alternative = "two.sided",conf.level = 0.99)

Exact binomial test	nature du test
data : 840 and 1000	données
number of successes=840, number of trials=1000, p-value=0.00135	p-value :probabilité de rejet
alternative hypothesis : true probability of success is not equal to 0.8	hypothèse alternative
99 percent confidence interval :	intervalle de confiance à 99%
0.8080225 0.8686768	
sample estimates :	estimations sur l'échantillon
probability of success 0.84	probabilité du succès

p-value = 0.00135 $<$ α = 0.01 $\Rightarrow$ rejet de H_0