
TD n° 4 : Estimation et tests (suite)

Exercice 1. (Expérience de Mendel) Reprenons l'étude de Mendel sur 557 petits pois.
La couleur est codée par un gène présentant deux allèles, C et c, correspondant aux couleurs jaune et vert. Le jaune est dominant, le vert est récessif. La forme est portée par un autre gène à deux allèles : R pour rond, dominant et r pour ridé, récessif.

Il y a donc 16 génotypes probables (les petits pois sont diploïdes), mais on n'observe que 4 phénotypes : jaune rond, jaune ridé, vert rond et vert ridé.

1) Lister les 16 génotypes et trouver, pour chacun d'eux, les phénotypes correspondants.
2) On suppose que les 16 génotypes sont équiprobables. Calculer les probabilités de chacun des phénotypes. En déduire les effectifs théoriques sur un échantillon de 557 petits pois.
3) Les effectifs observés par Mendel, donnés dans le tableau ci-dessous, diffèrent-ils des effectifs théoriques ? On donne pour aide numérique que $\mathbb{P}(\chi^2_3 > 0.45) \approx 0.93$. Discuter alors de votre choix.

Phénotype	jaune rond	jaune ridé	vert rond	vert ridé
Effectif	315	102	108	32

Correction:

1) Les 16 génotypes sont :

$$\begin{array}{c} CCRR - CCRr - CCrr - CCrR \\ CcRR - CcRr - Ccrr - CcRr \\ cCRR - cCRr - cCrr - cCrR \\ ccRR - ccRr - ccrr - ccrR \end{array}$$

Les 4 phénotypes sont alors :

$[CR]$	jaune et rond	$CCRR - CCRr - CCrR - CcRR - CcRr - CcRr - cCRR - cCRr - cCrR$
$[Cr]$	jaune et ridé	$CCrr - Ccrr - cCrr$
$[cR]$	vert et rond	$ccRR - ccRr - ccrR$
$[cr]$	vert et ridé	$ccrr$

2) Ainsi, on peut en déduire facilement la probabilité d'apparition de chacun des 4 phénotypes :

$$\mathbb{P}([CR]) = \frac{9}{16} \ , \quad \mathbb{P}([Cr]) = \frac{3}{16} \ , \quad \mathbb{P}([cR]) = \frac{3}{16} \ , \quad \mathbb{P}([cr]) = \frac{1}{16} \ .$$

Donc pour $n = 557$ petits pois, on s'attend à trouver les "effectifs théoriques" suivants :

$$\begin{aligned} n\mathbb{P}([CR]) &= 557 \times \frac{9}{16} \simeq 313 \text{ petits pois jaunes et ronds} \\ n\mathbb{P}([Cr]) &= 557 \times \frac{3}{16} \simeq 104.5 \text{ petits pois jaunes et ridés} \\ n\mathbb{P}([cR]) &= 557 \times \frac{3}{16} \simeq 104.5 \text{ petits pois verts et ronds} \\ n\mathbb{P}([cr]) &= 557 \times \frac{1}{16} \simeq 35 \text{ petits pois verts et ridés} \end{aligned}$$

3) Cela donne donc le tableau suivant :

Phénotype	jaune rond	jaune ridé	vert rond	vert ridé	total
Eff. obs.	315	102	108	32	557
Eff. théoriques	313	104.5	104.5	35	557
$\frac{(obs-th.)^2}{th}$	0.01	0.06	0.12	0.26	0.45

L'hypothèse nulle à tester $\mathcal{H}_0$ est : "Les phénotypes observés proviennent bien d'une loi mendélienne". La statistique de test du χ^2 vaut 0.45 et donc la probabilité d'observée une telle valeur rien que par hasard est donc de 0.93. On garde donc l'hypothèse nulle à 3 d.d.l.

```python
# Code Python
import numpy as np
from scipy.stats import chisquare,chi2
obs = np.array([315 , 102 , 108 , 32])
th = np.array([313 , 104.5 , 104.5 , 35])
chi2_stat, pval = chisquare(obs, th)
1 - chi2.cdf(chi2_stat, df=3)
# pval = 0.9303804770054722
# chi2_stat = 0.44695590268147856
```

Exercice 2.

Sur 100 individus issus de croisements de deux hétérozygotes Aa, on a observé 36 phénotypes $[a]$, récessifs. Le but de l'étude est de tester si AA est mortel.

a) Soit S le nombre de sujets de phénotype $[a]$, on suppose que les 4 génotypes sont équiprobables. Donner la loi de S.

b) Les observations suivent-elles une loi de Mendel classique ?

c) Écrire les hypothèses nulles et alternatives pour tester si AA est mortel.

d) Conclure sur les données observées.

On donne pour aide numérique que $\mathbb{P}(\chi^2_1 > 0.321) \approx 0.5716$ et $\mathbb{P}(\chi^2_1 > 6.453) \approx 0.0111$

Correction:

a) Les 4 génotypes sont :

$$aa - aA - Aa - AA$$

Les 2 phénotypes sont alors :

$$\begin{array}{c|l} [A] & AA - aA - Aa \\ [a] & aa \end{array}$$

Ainsi, on peut en déduire facilement la probabilité d'apparition de chacun des deux phénotypes :

$$\mathbb{P}([a]) = \frac{1}{4} \ , \quad \mathbb{P}([A]) = \frac{3}{4} \ ,$$

b) On a les hypothèses suivantes :
$\mathcal{H}_0$: Les observations suivent une loi de Mendel
vs.
$\mathcal{H}_1$: Les observations ne suivent pas une loi de Mendel.
Dans le cas de Mendel classique (où AA n'est pas mortel), on a ceci :

Phénotype	$[a]$	$[A]$	total
Eff. obs.	36	64	100
Eff. théoriques	$100 \times 1/4 = 25$	$100 \times 3/4 = 75$	100
$\frac{(obs-th.)^2}{th}$	4.84	1.613	6.453

$p-value \approx 0.0111 < 0.05$ donc on rejette $\mathcal{H}_0$ notre échantillon ne suit pas la répartition "normale" que devraient avoir les phénotypes.

c) $\mathcal{H}_0$: Les observations suivent la répartition théorique dans laquelle AA est mortel
vs.
$\mathcal{H}_1$: Les observations ne suivent pas la répartition théorique dans laquelle AA est mortel.

d) Dans le cas où AA est mortel, on n'a plus que 3 génotypes possibles : Aa, aA, aa. Donc la répartition des phénotypes devient la suivante : $\mathbb{P}([A]) = 2/3$ et $\mathbb{P}([a]) = 1/3$. On a alors :

Phénotype	$[a]$	$[A]$	total
Eff. obs.	36	64	100
Eff. théoriques	$100 \times 1/3 \approx 33.33$	$100 \times 2/3 \approx 66.67$	100
$\frac{(obs-th.)^2}{th}$	0.214	0.107	0.321

$p-value \approx 0.5716 > 0.05$ donc on ne rejette pas $\mathcal{H}_0$ notre échantillon suit la répartition théorique dans laquelle AA est mortel.

Exercice 3. On dispose d'un jeu de données composé de deux échantillons supposés gaussiens indépendants, dont le résumé est

Échantillon 1	Échantillon 2
$n_1 = 55$	$n_2 = 60$
$\bar{x}_1 = 18.4$	$\bar{x}_2 = 16.5$
$s_1 = 8.6$	$s_2 = 13.7$

Supposons que σ_1 et σ_2 sont connus tels que $\sigma_1 = 8.7$ et $\sigma_2 = 13.8$

1) Construire un intervalle de confiance pour la différence des moyennes au niveau 98%.
2) Tester $\mathcal{H}_0$: $\mu_1 - \mu_2 = 2.5$ contre $\mathcal{H}_1$: $\mu_1 - \mu_2 \neq 2.5$ avec $\alpha = .02$.
3) Tester $\mathcal{H}_0$: $\mu_1 - \mu_2 = 2.5$ contre $\mathcal{H}_1$: $\mu_1 - \mu_2 < 2.5$ avec $\alpha = .02$.
4) Refaire l'exercice en supposant σ_1 et σ_2 inconnus.

On donne $q_{0.99} = 2.3263$, $q_{0.98} = 2.0537$, $t_{0.99}(54) = 2.3974$, $t_{0.98}(54) = 2.1045$.

Correction:

1) **Cas 1, σ_1 et σ_2 connus :** Sans hypothèse d'égalité entre σ_1 et σ_2, Un IC au niveau $(1-\alpha)$ pour $\mu_1 - \mu_2$ est

$$\left(\bar{x}_1 - \bar{x}_2\right) \pm q_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, \quad \text{où}$$

$q_{1-\alpha/2}$ est le quantile d'ordre $(1-\alpha/2)$ d'une loi normale centrée réduite. Application numérique : $q_{1-\alpha/2} = q_{0.99} = 2.3263$:

$$\left(18.4 - 16.5\right) \pm 2.3263\sqrt{\frac{8.7^2}{55} + \frac{13.8^2}{60}} = [-3.06\,;\, 6.86]$$

Cas 2, σ_1 et σ_2 inconnus : Sans hypothèse d'égalité entre σ_1 et σ_2, Un IC au niveau $(1-\alpha)$ pour $\mu_1 - \mu_2$ est

$$\left(\bar{x}_1 - \bar{x}_2\right) \pm t_*\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, \quad \text{où}$$

t_* est le quantile d'ordre $(1-\alpha/2)$ de $t(k)$, où k est le plus petit nombre entre $(n_1 - 1)$ et $(n_2 - 1)$. Application numérique : $k = 54$ et $t_* = t_{0.99}(54) = 2.3974$:

$$\left(18.4 - 16.5\right) \pm 2.3974\sqrt{\frac{8.6^2}{55} + \frac{13.7^2}{60}} = [-3.17\,;\, 6.97]$$

Cas 1, σ_1 et σ_2 connus : Tester $\mathcal{H}_0 : \mu_1 - \mu_2 = 2.5$ contre $\mathcal{H}_1 : \mu_1 - \mu_2 \neq 2.5$ avec $\alpha = .02$. La statistique de test vaut :

$$T = \frac{(18.4 - 16.5 - 2.5)}{\sqrt{8.7^2/55 + 13.8^2/60}} = -0.2813$$

Le test est bilatéral donc il y a rejet de $\mathcal{H}_0$ si $|T| > q_{1-\alpha/2} = q_{0.99} = 2.3263$. Ici, on ne rejette pas $\mathcal{H}_0$ au niveau de 98%.

Cas 2, σ_1 et σ_2 inconnus : Tester $\mathcal{H}_0 : \mu_1 - \mu_2 = 2.5$ contre $\mathcal{H}_1 : \mu_1 - \mu_2 \neq 2.5$ avec $\alpha = .02$. La statistique de test vaut :

$$T = \frac{(18.4 - 16.5 - 2.5)}{\sqrt{8.6^2/55 + 13.7^2/60}} = -0.2837$$

Le test est bilatéral donc il y a rejet de $\mathcal{H}_0$ si $|T| > t_* = t_{0.99}(54) = 2.3974$. Ici, on ne rejette pas $\mathcal{H}_0$ au niveau de 98%.

Cas 1, σ_1 et σ_2 connus : Tester $\mathcal{H}_0 : \mu_1 - \mu_2 = 2.5$ contre $\mathcal{H}_1 : \mu_1 - \mu_2 < 2.5$ avec $\alpha = .02$. La statistique de test est inchangée $T = -0.2813$. C'est la règle de décision qui change, on veut faire un test unilatéral avec l'hypothèse alternative $\mathcal{H}_1 : \mu_1 - \mu_2 < 2.5$. Ici, on rejettera $\mathcal{H}_0$ si $T < q_\alpha = q_{0.02} = -2.0537$. Donc on garde $\mathcal{H}_0$ au niveau de 98%.

Cas 2, σ_1 et σ_2 inconnus : Tester $\mathcal{H}_0 : \mu_1 - \mu_2 = 2.5$ contre $\mathcal{H}_1 : \mu_1 - \mu_2 < 2.5$ avec $\alpha = .02$. La statistique de test est inchangée $T = -0.2837$. C'est la règle de décision qui change, on veut faire un test unilatéral avec l'hypothèse alternative $\mathcal{H}_1 : \mu_1 - \mu_2 < 2.5$. Ici, on rejettera $\mathcal{H}_0$ si $T < t_* = t_{0.02}(54) = -2.1045$. Donc on garde $\mathcal{H}_0$ au niveau de 98%.

Exercice 4. Une expérience de comparaison de deux traitements de plantes repose sur l'étude d'un échantillon de 88 plantes. Dans ce groupe, 40 plantes ont été sélectionnées au hasard pour subir le traitement 1 et les 48 plantes restantes ont subi le traitement 2. La moyenne et l'écart-type du poids des plantes dans les deux échantillons sont

	Traitement 1	Traitement 2
Moyenne	16.21	27.84
Écart-type	2.88	4.32

1) Donner un intervalle de confiance, au niveau 95%, de la différence des moyennes.
2) Le traitement 1 est un placebo, le traitement 2 est un nouvel insecticide. Cet insecticide ne sera mis sur le marché que si le poids des plantes ayant subi ce traitement est d'au moins 10 unités plus élevé qu'en l'absence de traitement. Écrire les hypothèses nulles et alternatives.
3) Quelle est la statistique de test ? Quelle est la région de rejet pour $\alpha = 0.05$?
4) Effectuer le test. Trouver la p-valeur et commenter.

On donne $q_{0.95} \approx 1.645$, $q_{0.975} \approx 1.96$, $t_{0.975}(39) \simeq 2.023$, $t_{0.95}(39) \approx 1.685$.

Correction:

1) Sans hypothèse d'égalité entre σ_1 et σ_2, un IC au niveau $(1-\alpha)$ pour $\mu_1 - \mu_2$ est

$$(\bar{x}_1 - \bar{x}_2) \pm t_* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, \quad \text{où}$$

t_* est le quantile d'ordre $(1-\alpha/2)$ de $t(k)$, où k est le plus petit nombre entre $(n_1 - 1)$ et $(n_2 - 1)$. Application numérique : $k = \min(39, 47)$ et $t_* = t_{0.975}(39) \simeq= 2.023$:

$$(16.21 - 27.84) \pm 2.02 \sqrt{\frac{2.88^2}{40} + \frac{4.32^2}{48}} = [-13.19\,;\, -10.07]$$

2) $\mathcal{H}_0 : \mu_1 - \mu_2 \geq \delta_0$ vs $\mathcal{H}_1 : \mu_1 - \mu_2 \ < \delta_0$: test unilatéral à gauche. Ici, $\delta_0 = -10$, la statistique de test est donc :

$$T = \frac{\bar{x}_1 - \bar{x}_2 + 10}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Application numérique :

$$T_{obs} = \frac{16.21 - 27.84 + 10}{\sqrt{\frac{2.88^2}{40} + \frac{4.32^2}{48}}} = -2.11$$

et $t_{**} = t_{0.05}(39) \simeq -1.6849$.

3) La région de rejet est $]-\infty, -1.6849]$ pour un risque de 5%.

4) Comme $T_{obs} < -1.6849$, on rejette $\mathcal{H}_0$. Donc le nouvel insecticide semble augmenter le poids de plus de 10 unités.

Exercice 5. Un chercheur en pharmacie veut déterminer si le médicament qu'il a développé a comme effet secondaire de faire baisser la pression artérielle. L'étude commence par l'enregistrement de la pression chez 15 jeunes étudiantes avant le traitement. Ensuite, ces 15 femmes ont pris ce médicament quotidiennement pendant 6 mois et leur pression artérielle a été mesurée à la fin de cette période.

Les données sont

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Avant	70	80	72	76	76	76	72	78	82	64	74	92	74	68	84
Après	68	72	62	70	58	66	68	52	64	72	74	60	74	72	74

1) S'agit-il d'un échantillon apparié ou indépendant ?

2) Quelle hypothèse nulle et alternative doit-on poser ?

3) Calculer un intervalle de confiance au niveau 95% de la différence de pression artérielle.

4) Les données montrent-elles une baisse significative de la pression ?

Correction:

1) échantillon apparié. Donc ici $n = 15$.

2) On va poser comme hypothèse :

$\mathcal{H}_0$: "Le médicament ne fait pas baisser la tension" v.s "Le médicament fait baisser la tension "

N.B : Dans les études pharmaceutiques, on fait toujours l'hypothèse nulle $\mathcal{H}_0$: "Le médicament n'a pas d'effet" car ainsi on contrôle le risque de première espèce $\alpha = P\Big(\text{Rejeter } \mathcal{H}_0 | \mathcal{H}_0 \text{ vraie}\Big)$ qui est le plus risqué pour le patient. Ici, $\alpha = P\Big(\text{Dire que le médicament fait baisser la tension alors qu'il n'a pas d'effet.}\Big)$ On traduit donc par :

$$\mathcal{H}_0 : \mu_1 \leq \mu_2 \text{ v.s } \mathcal{H}_1 : \mu_1 > \mu_2$$

il s'agit d'un test unilatéral à droite donc la zone de rejet (pour échantillon apparié) de $\mathcal{H}_0$ est de la forme :

$$[t_{1-\alpha}(n-1); +\infty[= [t_{0.95}(15-1), +\infty[= [1.7613, +\infty[$$

3) On calcule les différences D :

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Avant X_i	70	80	72	76	76	76	72	78	82	64	74	92	74	68	84
Après Y_i	68	72	62	70	58	66	68	52	64	72	74	60	74	72	74
Diff. $D = X_i - Y_i$	2	8	10	6	18	10	4	26	18	-8	0	32	0	-4	10

On calcule la moyenne $\bar{D} = 8.8$ et $S_D^2 = \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \bar{D})^2 = 120.4571$. On en déduit l'intervalle de confiance de niveau 95% :

$$\bar{D} \pm t_* \frac{S_D}{\sqrt{n}}$$

Application numérique :

$$8.8 \pm t_{0.975}(14) \times \frac{\sqrt{120.4571}}{15}$$

soit avec $t_{0.975}(14) = 2.1448$, on obtient $[2.722\,;\,14.878]$.

4) Pour le test unilatéral à droite de $\mathcal{H}_0 : \mu_1 \leq \mu_2$ on calcule la statistique de test pour échantillon apparié :

$$T = \frac{\bar{D} - \delta_0}{S_D/\sqrt{n}} \quad \text{ici } \delta_0 = 0,$$

Et, $T_{obs} = \dfrac{8.8}{\sqrt{120.4571/15}} \simeq 3.1$ qui appartient à la zone de rejet de $\mathcal{H}_0$: $[1.7613, +\infty[$ pour un risque de 5%. Donc le médicament fait bien baisser la tension des patients.