# Tests et régions de confiance

v1.3 mars 2019 / non relue sérieusement

Ces notes présentent les aspects les plus élémentaires des tests d'hypothèses simples et des régions de confiance dans un des cadres les plus simples : celui de l'estimation d'une proportion. Un accent est mis sur le fait que tout peut être fait de manière exacte, sans approximations liées à des théorèmes limites.

## 1 Un problème concret

Une population est composée d'individus de deux types : le type $X$ et le type $Y$. On note $p$ la proportion d'individus de type $X$. On cherche à estimer $p$ sans avoir à observer toute la population.

On suppose (c'est une idéalisation) que l'on est capable de répéter $n$ fois de manière indépendante l'action suivante : tirer au sort de manière uniforme un individu et observer son type. Une fois ces $n$ actions effectuées, on dispose du nombre $n_X$ de fois qu'un individu de type $X$ a été tiré au sort. Il est naturel d'espérer (si $n$ est grand) que $n_X/n$ est proche de $p$. Au vu de $n_X/n$ on peut, par exemple, se demander si $p$ est proche de $1/2$ ou non. C'est une question vague dont on donnera un sens précis dans la suite.

## 2 Quelques remarques informelles

1. On tire 10000 fois au sort. La question « quelle est la probabilité que $p$ vaille $1/2$ (ou celle que $p$ soit compris entre 0.45 et 0.55) sachant que l'on a obtenu plus de 60% de $X$ » est *a priori* stupide : cette probabilité vaut 1 si $p = 1/2$ et 0 sinon (à moins de modéliser le problème avec un $p$ aléatoire...).
2. On tire 10000 fois au sort. La question « si $p = 1/2$, quelle est la probabilité que l'on obtienne plus de 60% de $X$ » a un sens naturel.
3. Si $p = 1/2$, obtenir plus de 60% de $X$ tirant 10 fois au sort n'a rien d'improbable.
4. Si $p = 1/2$, obtenir plus de 60% de $X$ en tirant 10000 fois au sort est très improbable.
5. Si on tire 10 fois au sort et si $p = 1/2$, la probabilité de n'obtenir que des $X$ est $1/2^{10} = 1/1024$ : c'est très improbable.
6. Si on tire 10 fois au sort et si $p = 1/2$, la probabilité d'obtenir par exemple la succession de résultats $XYXXYYYXYX$ est également $1/2^{10} = 1/1024$ : c'est tout aussi improbable.

## 3 Une modélisation

Soit $p \in [0, 1]$. Soit $(B_i)_{i\geq 1}$ une suite de variables aléatoires indépendantes de Bernoulli de paramètre $p$. Pour marquer la dépendance en $p$ on notera $\mathbb{P}_p$ la mesure de probabilité (on peut si on le souhaite construire toutes ces variables aléatoires sur un même espace probabilisé et considérer différentes mesures de probabilité). Soit $n \geq 1$ un entier. On modélise la quantité $n_X/n$ de la section précédente par la variable aléatoire

$$F = F_n = \frac{1}{n}\sum_{i=1}^{n} B_i.$$

# 4 Test de l'hypothèse $p = 1/2$ contre l'hypothèse $p \neq 1/2$

**Définitions dans ce cadre.**

— Un test de l'hypothèse $p = 1/2$ contre l'hypothèse $p \neq 1/2$ au seuil $0 < \alpha < 1$ est une application

$$\text{test} : \{0,1\}^n \to \{\text{acceptation}, \text{rejet}\}$$

telle que

$$\mathbb{P}_{1/2}\big[\text{test}(B_1, \ldots, B_n) = \text{rejet}\big] \leq \alpha. \tag{1}$$

En mots :

1. La décision de rejeter (ou non) l'hypothèse $p = 1/2$ au profit de l'hypothèse $p \neq 1/2$ ne dépend que des variables aléatoires $B_1, \ldots, B_n$.
2. La probabilité de rejeter à tort est plus $\alpha$.

— La puissance d'un test est l'application $\pi : [0,1] \setminus \{1/2\} \to [0,1]$ définie par

$$\pi(p) = \mathbb{P}_p\big[\text{test}(B_1, \ldots, B_n) = \text{rejet}\big].$$

En mots : c'est la probabilité de rejeter à raison l'hypothèse $p = 1/2$. Cette probabilité dépend de la valeur de $p$. On préfère bien sûr les tests puissants.

**Résultats corrects et incorrects.**

1. Le résultat d'un test est correct dans deux cas :
   (a) Lorsque $p = 1/2$ et $\text{test}(B_1, \ldots, B_n) = \text{acceptation}$.
   (b) Lorsque $p \neq 1/2$ et $\text{test}(B_1, \ldots, B_n) = \text{rejet}$.
2. Le résultat d'un test est incorrect dans deux cas :
   (a) Lorsque $p = 1/2$ et $\text{test}(B_1, \ldots, B_n) = \text{rejet}$. La probabilité est majorée par le seuil.
   (b) Lorsque $p \neq 1/2$ et $\text{test}(B_1, \ldots, B_n) = \text{acceptation}$. La probabilité est majorée par $1 - \pi(p)$.

**Exemple de test trivial.** Le test défini par

$$\forall (b_1, \ldots, b_n) \in \{0,1\}^n,\ \text{test}(b_1, \ldots, b_n) = \text{acceptation}$$

est un test pour tout seuil $\alpha > 0$. Sa puissance est nulle et ce test est bien entendu sans intérêt.

**Vocabulaire.** Lorsqu'un test est défini, on dit que « l'hypothèse $p = 1/2$ est rejetée » lorsque l'évènement $\{\text{test}(B_1, \ldots, B_n)\}$ a lieu et que « l'hypothèse $p = 1/2$ n'est pas rejetée » sinon.

**Mise en place d'un test dans ce cadre.** De manière vague et au moins pour $n$ grand, le comportement est le suivant :

— Si $p$ est proche de $1/2$ alors, avec une grande probabilité, $F$ est proche de $1/2$.
— Sinon, avec une grande probabilité, $F$ n'est pas proche de $1/2$.

Cela conduit à définir le test suivant, où $r$ est un réel à définir : pour tout $(b_1, \ldots, b_n) \in \{0,1\}^n$,

$$\text{test}(b_1, \ldots, b_n) = \text{acceptation si } \frac{b_1 + \cdots + b_n}{n} \in [1/2 - r, 1/2 + r] \text{ et } \text{test}(b_1, \ldots, b_n) = \text{rejet sinon.}$$

La condition (1) s'écrit ici :

$$\mathbb{P}_{1/2}(F \notin [1/2 - r, 1/2 + r]) \leq \alpha.$$

Comme on souhaite un test le plus puissant possible, on chosit le plus petit réel $r \geq 0$ vérifiant la condition précédente. Ce réel $r$ peut-être explicité (de manière exacte ou approchée) car la loi de $F$ lorsque $p = 1/2$ est connue !

Posons

$$A = [1/2 - r, 1/2 + r].$$

On appelle $A$ « l'intervalle d'acceptation ».

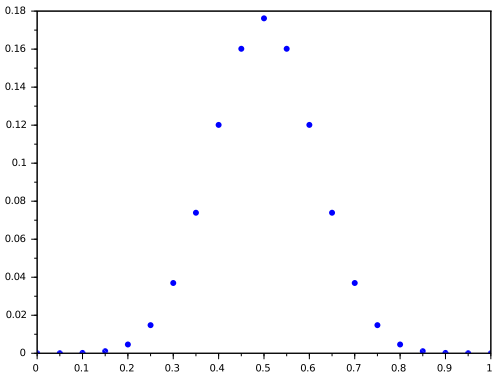


FIGURE 1 – Loi de $F_{20}$ sous $P_{1/2}$. Pour $\alpha = 0.05$ et $n = 20$ on a $r = 4/20$.

**Protocole concret associé au test.** Le protocole du test est alors le suivant :

— Si $n_A/n$ n'appartient pas à l'intervalle $A$ alors on rejette l'hypothèse $p = 1/2$.
— Sinon on ne rejette pas l'hypothèse $p = 1/2$.

Le sens précis de « on rejette l'hypothèse $p = 1/2$ » n'est rien d'autre que celui décrit dans le protocole précédent : cela signifie que $n_A/n$ n'appartient pas à l'intervalle défini précédemment ! De manière vague on peut dire la chose suivante : si le test amène à rejeter l'hypothèse $p = 1/2$ alors ou l'hypothèse est effectivement fausse ou, lors du tirage au sort, il s'est passé un évènement de probabilité au plus $\alpha$. Mais cette formulation est trop vague pour avoir un sens. Il peut être utile de réfléchir à nouveau sur les remarques de la section 2.

**Risques de première et de seconde espère.** Considérons les différents risques.

1. Le risque de première espèce est la probabilité de rejeter à tort l'hypothèse $p = 1/2$. Par définition, c'est au plus égal au seuil $\alpha$ du test.
2. Le risque de seconde espèce dépend de la valeur de $p$. C'est

$$\mathbb{P}_p(F \in A).$$

On espère que ce risque est petit au moins pour $p$ suffisamment loin de $1/2$ et pour $n$ suffisamment grand. La puissance du test le complément à 1 de ce risque :

$$\pi(p) = 1 - \mathbb{P}_p(F \in A) = \mathbb{P}_p(F \notin A).$$

On espère que la puissance du test est proche de 1, au moins pour $p$ suffisamment loin de $1/2$ et pour $n$ suffisamment grand. On a par contre $\pi(1/2) \leq \alpha$. Pour tout $n$ et pour tout $p$ suffisamment proche de $1/2$, la puissance du test est petite. C'est inévitable : on ne peut espérer distinguer $p = 1/2$ de $p = 1/2 + 1/10000$ dans nos observations que si $n$ est très grand.

**Quelques remarques et questions.**

— Le réel $r$ ne dépend pas de la taille de la population dans notre problème concret. Il dépend par contre de $n$, i.e. du nombre d'individus sondés.
— Quel est le comportement asymptotique de $r$ lorsque $n$ tend vers l'infini ?
— Comment se comporte la puissance si on diminue le seuil ? De manière plus caricaturale, que dire des risques de première et seconde espèces lorsque l'on rejette toujours ? Ou au contraire lorsque l'on ne rejette jamais ?
— Peut-on trouver, pour un seuil $\alpha$ fixé, un test qui optimise la puissance $\pi(p)$ pour tout $p \neq 1/2$ ?
— Quel pourrait être (définition et exemple) un test de l'hypothèse $p \geq 1/2$ contre l'hypothèse alternative $p < 1/2$ ?

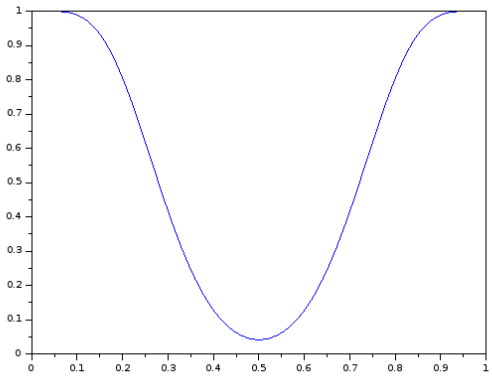


FIGURE 2 – La puissance $\pi$ en fonction de $p$ pour $n = 20$ et $\alpha = 0.05$.

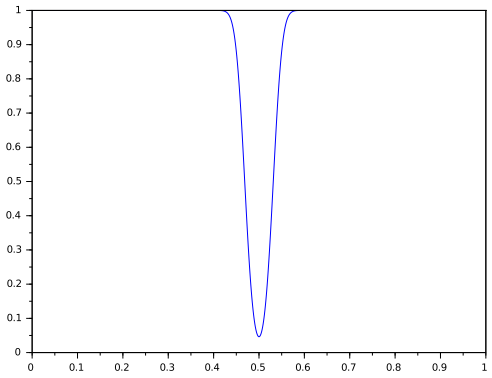


FIGURE 3 – La puissance $\pi$ en fonction de $p$ pour $n = 1000$ et $\alpha = 0.05$.

## 5 Test de de l'hypothèse $p = q$ contre l'hypothèse $p \neq q$

Soit $q \in [0, 1]$. On peut généraliser (définitions, idées etc.) ce qui a été fait précédemment en remplaçant $1/2$ par $q$. On obtient le protocole suivant pour le test de l'hypothèse $p = q$ contre l'hypothèse alternative $p \neq q$. On considère le plus petit réel $r(q) \geq 0$ tel que

$$\mathbb{P}_q(F \notin [q - r(q), q + r(q)]) \leq \alpha.$$

Notons

$$A(q) = [q - r(q), q + r(q)] \cap [0, 1]$$

l'intervalle d'acceptation.

— Si $n_A/n$ n'appartient pas à l'intervalle d'acceptation $A(q)$ alors on rejette l'hypothèse $p = q$.
— Sinon on ne rejette pas l'hypothèse $p = q$.

On a choisi ici encore un intervalle symétrique autour de $q$. C'est un choix arbitraire et contestable. Mais le test qui précède rempli cependant le cahier des charges. On pourrait par exemple prendre pour intervalle d'acceptation un intervalle minimal (pour la longueur) parmi tous les intervalles $A$ tels que $\mathbb{P}_q(F \in A) \geq 1 - \alpha$. On pourrait également (est-ce très différent ?) poser

$$A(q) = [G(q), D(q)] \tag{2}$$

où

$$G(q) = \min\{f \in \{0, 1/n, \ldots, 1\} : \mathbb{P}_q(F = f) \geq \eta\}$$

et

$$D(q) = \max\{f \in \{0, 1/n, \ldots, 1\} : P_q(F = f) \geq \ \eta\}$$

avec $\eta$ maximal parmi les $\eta$ tels que l'intervalle $A(q)$ ainsi associé vérifie $\mathbb{P}_q(A(q)) \geq 1-\alpha$. C'est l'intervalle choisi plus bas pour les illustrations.

# 6 Région de confiance

Dans ce qui précède on a par exemple testé l'hypothèse $p = 1/2$ contre l'hypothèse alternative $p \neq 1/2$. On a déjà noté que, pour un entier $n$ donné, on ne pouvait espérer voir sur nos données une différence suivant si $p$ vaut $1/2$ ou si $p$ est proche de $1/2$. C'est une des raisons qui peut amener à considérer une région de confiance aléatoire $C$ pour $p$ qui peut être défini concrètement ainsi (une fois que l'on a défini un test pour toute valeur de $q$) :

$$C(F) = C = \{q \in [0, 1] : \text{ pas de rejet si on teste l'hypothèse } p = q \text{ contre l'hypothèse } p \neq q\}.$$

Il est important de noter que $C$ est aléatoire : il dépend de la variable aléatoire $F$. Cela se voit peut-être mieux en explicitant $C$ ainsi :

$$C = \{q \in [0, 1] : F \in A(q)\}.$$

La région $C$ vérifie la propriété fondamentale suivante :

$$\forall p \in [0, 1],\ \mathbb{P}_p(p \in C) \geq 1 - \alpha. \tag{3}$$

En effet, pour tout $p \in [0, 1]$,

$$\begin{aligned} \mathbb{P}_p(p \in C) &= \mathbb{P}_p(\text{non rejet de l'hypothèse « le paramètre vaut } p \text{ »}) \\ &= 1 - \mathbb{P}_p(\text{rejet de l'hypothèse « le paramètre vaut } p \text{ »}) \\ &\geq 1 - \alpha. \end{aligned}$$

On a choisi dans cette présentation de commencer par construire des tests et d'utiliser ceux-ci pour construire une région de confiance. On peut procéder dans l'autre sens (c'est ce que l'on trouve dans beaucoup de documents sur le sujet) : définir une région de confiance comme étant une région (souvent un intervalle) aléatoire (dépendant de $F$) contenant le paramètre étudié (ici $p$) avec probabilité au moins $1 - \alpha$ ; construire une telle région de confiance ; définir ensuite des tests à partir de cette région de confiance (on rejette l'hypothèse « le paramètre vaut $q$ » si $q$ n'appartient pas à $C$).

# 7 Intervalles d'acceptation et régions de confiance : une vision plus géométrique

Posons

$$\mathcal{G} = \{(f, p) \in [0, 1] \times [0, 1] : f \in A(p)\}.$$

Pour $p \in [0, 1]$ donné, on a

$$A(p) = \{f \in [0, 1] : (f, p) \in \mathcal{G}\}$$

et, pour tout $f \in [0, 1]$,

$$C(f) = \{p \in [0, 1] : (f, p) \in \mathcal{G}\}.$$

# 8 Résultats asymptotiques

La loi des grands nombres et le théorème central limite nous renseignent sur le comportement asymptotique (en $n$) des intervalles d'acceptation et des régions de confiance.

Faisons plutôt un peu de cuisine (comme dans beaucoup de documents sur le sujet). C'est un point de vue pragmatique.

Pour tout $p \in [0,1]$ et tout $n$ assez grand on a

$$\sqrt{n}\frac{F_n - p}{\sqrt{p(1-p)}} \approx G \text{ en loi sous } \mathbb{P}_p$$

et donc (en faisant comme si c'était une égalité en loi ; c'est là que commence la cuisine)

$$\mathbb{P}_p\left(\sqrt{n}\left|\frac{F_n - p}{\sqrt{p(1-p)}}\right| \leq 1.96\right) \approx 0.95.$$

Pour le seuil $\alpha = 0.05$ on peut donc prendre :

$$A_n(p) = \left[p - 1.96\sqrt{p(1-p)/n}, p + 1.96\sqrt{p(1-p)/n}\right].$$

On en déduit des régions de confiance un peu désagréables.

On a aussi, pour tout $p \in [0,1]$ et tout $n$ assez grand,

$$\sqrt{n}\frac{F_n - p}{\sqrt{F_n(1-F_n)}} \approx G \text{ en loi sous } \mathbb{P}_p$$

et donc

$$\mathbb{P}_p\left(\sqrt{n}\left|\frac{F_n - p}{\sqrt{F_n(1-F_n)}}\right| \leq 1.96\right) \approx 0.95.$$

On en déduit, pour $\alpha = 0.05$, des zones de rejet un peu désagréables et des régions de confiance plus sympathiques :

$$C(f) = \left[f - 1.96\sqrt{f(1-f)/n}, f + 1.96\sqrt{f(1-f)/n}\right]. \tag{4}$$

Dans les figures qui suivent est représentée, pour différentes valeurs de $n$, la partie $\mathcal{G}$ (voir section 7) calculée de deux manières : de manière exacte avec (2) et de manière cuisinée avec (4).

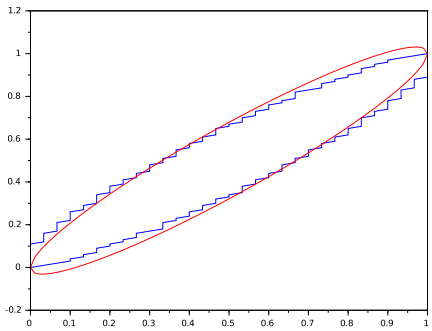


FIGURE 4 – $\mathcal{G}$ pour $n = 30$ : exact en bleu ; cuisiné en rouge

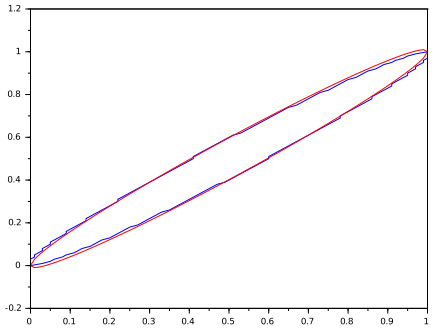


FIGURE 5 – $\mathcal{G}$ pour $n = 100$ : exact en bleu ; cuisiné en rouge

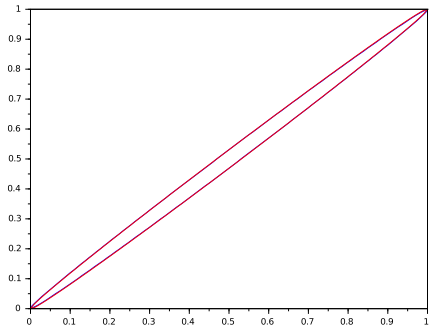


FIGURE 6 – $\mathcal{G}$ pour $n = 1000$ : exact en bleu ; cuisiné en rouge

# 9 Quelques questions supplémentaires

1. Mille équipes testent indépendamment une hypothèse avec un seuil de 0.05. On suppose que l'hypothèse est effectivement vérifiée. Que dire du nombre d'équipes rejetant l'hypothèse ?
2. Et si les mille équipes ne testent pas la même hypothèse ?
3. Reproduire (par exemple en python) les différents graphiques de ce document.
4. Quelle est la nature géométrique de

$$\{t + 1.96\sqrt{t(1-t)/n}, t \in [0,1]\} \cup \{t - 1.96\sqrt{t(1-t)/n}, t \in [0,1]\}$$

pour un entier $n$ fixé ?