TD Test, test multiple

Philippe Veber

11 février 2026

Les exercices comportent des indications pour le langage R, mais peuvent bien entendu être réalisés dans le langage de votre choix.

1 p-valeur, taille d'effet, taille d'échantillon

Les exercices de cette première partie reviennent sur la notion de test statistique au sens introduit par Fisher. Dans ce cadre, on considère une hypothèse, appelée *hypothèse nulle* et notée $\mathcal{H}_0$, que l'on cherche à *réfuter* à partir d'observations expérimentales. La démarche consiste à :

1. collecter les données d'une expérience
2. calculer un résumé de ces données appelé *statistique de test*
3. déterminer la probabilité p d'observer, *si on répétait infiniment l'expérience dans des conditions où $\mathcal{H}_0$ est vraie*, une valeur plus grande de la statistique de test que celle observée dans l'expérience que l'on a réellement menée.

La probabilité p est appelée p-valeur. Si on trouve une valeur faible pour p cela signifie :

— ou bien que $\mathcal{H}_0$ est vraie et que l'on a observé un évènement rare,

— ou bien que $\mathcal{H}_0$ est fausse.

Plus la valeur de p est faible, plus on jugera que le niveau de preuve *contre* $\mathcal{H}_0$ est élevé. Cette démarche de test est souvent appelée *significance test* dans la littérature, et est celle que l'on pratique le plus souvent en sciences expérimentales. Attention, il existe d'autres approches pour le test d'hypothèse (test de Neyman-Pearson et test bayésien notamment) que nous n'aborderons pas ici.

Exercice 1
L'usine Duclou vient de recevoir une nouvelle machine censée produire des vis de longueur 10 mm avec un écart-type de 0.1 mm. Les ouvri·er·ères décident de la tester avant de la mettre en production, et produisent pour cela 10 vis. La longueur moyenne trouvée dans cet échantillon est de 10.098. À l'aide d'une simulation, calculez une grandeur mesurant à quel point ce résultat est plausible si la machine a effectivement les caractéristiques annoncées par le constructeur.

Exercice 2
On considère un dé équilibré à 6 faces.

1. Calculez par simulation la probabilité avec laquelle n lancers du dé donnent au moins k fois la face 6. Utilisez pour cela la fonction `sample`.

2. Avec votre fonction, calculez la probabilité que parmi 60 lancers 18 ou plus tombent sur la face 6.
3. Déterminez une formule pour cette probabilité en fonction de k et n, et vérifiez le résultat obtenu par simulation.
4. Quel nom donne-t-on à ce genre de probabilité ? Donnez-en une définition générale.

Exercice 3
Nous cherchons ici à implémenter le test t (du moins sa variante la plus simple). Soit un échantillon iid (indépendant et identiquement distribué) $X_1, \ldots, X_n$ tiré selon une loi gaussienne de moyenne μ et de variance σ^2 (on suppose σ connu). On souhaite tester contre l'hypothèse nulle $H_0 : \mu = \mu_0$ pour un μ_0 donné.

1. Soient $Z \sim N(\mu_Z, \sigma^2)$ et $\alpha > 0$. Quelle est la distribution de αZ ?
2. Soient $Z \sim N(\mu_Z, \sigma^2)$ et $W \sim N(\mu_W, \sigma^2)$ deux variables aléatoires indépendantes. Quelle est la distribution de $Z + W$? Si on a $\mu_Z = \mu_W$, $Z + W$ et $2Z$ sont-elles identiquement distribuées ? Pourquoi ?
3. Quelle est la distribution de la moyenne de l'échantillon $X_1, \ldots, X_n$?
4. Proposez un résumé statistique pour l'échantillon dont la distribution soit $N(0, 1)$ sous H_0.
5. Utilisez la fonction `pnorm` pour implémenter un test contre H_0. Comparez les résultats de votre fonction à ceux de `t.test`. Explication ?

Exercice 4 (Distribution de p)
On l'oublierait facilement, mais la p-valeur est une variable aléatoire, au même titre qu'une mesure expérimentale ou la moyenne d'un échantillon, puisque sa valeur dépend d'observations considérées comme aléatoires. On peut donc très légitimement se poser la question de sa distribution de probabilité. C'est ce que nous allons investiguer empiriquement dans cet exercice, en représentant graphiquement la distribution de la p-valeur lorsque l'on répète un test sur des données générées aléatoirement.

1. écrivez une fonction qui génère un échantillon tiré selon une distribution gaussienne centrée réduite (moyenne 0 et variance de 1), effectue un test t sur cet échantillon contre l'hypothèse nulle $\mu = 0$, et retourne la p-valeur obtenue. Construisez à l'aide de cette fonction un échantillon de p-valeurs.
2. affichez l'histogramme des p-valeurs obtenues
3. formulez une hypothèse sur la distribution des p-valeurs sous l'hypothèse nulle
4. quelle en est la conséquence pratique ?
5. recommencez en tirant les échantillons d'une gaussienne réduite de moyenne à 1 et refaîtes le test contre l'hypothèse nulle $\mu = 0$. De quelle manière la distribution change-t-elle ? Est-ce attendu ? Quelle loi permettrait de modéliser cette nouvelle distribution ?

Exercice 5 (Taille d'effet et p-valeur)
Nous allons ici étudier la notion de taille d'effet et son influence sur la distribution de la valeur p. Très typiquement dans une expérience on cherche à déterminer comment un facteur expérimental X influence une mesure Y. On appelle *taille d'effet* la variation de Y quand on fait varier X. Par exemple, on peut s'intéresser à l'expression d'un gène chez une espèce et à sa variation quand on mute le génome de ladite espèce. Dans ce cas, la taille d'effet est la différence d'expression entre les génotypes sauvage et mutant.

1. écrivez une fonction de simulation prenant un argument θ qui génère un échantillon de taille 30 selon une loi gaussienne réduite de moyenne θ, et retourne la p-valeur du test t contre l'hypothèse nulle d'une moyenne à 0.
2. utilisez votre fonction pour construire un `data.frame` contenant pour chaque ligne le résultat d'une simulation, avec en première colonne la taille d'effet et en deuxième colonne la valeur p obtenue. Le `data.frame` devra contenir un nombre suffisant de répétitions pour chaque taille d'effet.
3. représentez à l'aide d'un *boxplot* la façon dont la distribution des p-valeurs varie avec la valeur de θ
4. interprétation ?
5. mêmes questions en fixant θ à 1 et en augmentant la variance de la loi gaussienne.

Exercice 6 (Taille d'échantillon et p-valeur)
On s'intéresse ici à l'influence de la taille de l'échantillon utilisé dans un test sur la valeur de la p-valeur.

1. écrivez une fonction qui à une taille d'échantillon n associe une valeur p obtenue en générant un échantillon de taille n dans une loi gaussienne réduite de moyenne 1, et en réalisant un test t contre l'hypothèse nulle d'une loi gaussienne centrée réduite.
2. utilisez votre fonction pour construire un `data.frame` contenant pour chaque ligne le résultat d'une simulation, avec en première colonne la taille de l'échantillon et en deuxième colonne la p-valeur obtenue. Le `data.frame` devra contenir un nombre suffisant de répétitions pour chaque taille d'échantillon.
3. Visualisez votre tableau de simulations avec un *boxplot*, pour mettre en évidence la façon dont la distribution de valeurs p évolue avec la taille de l'échantillon.
4. Interprétation ?

Exercice 7
On considère deux populations d'insectes dont on mesure la taille. Est-il possible de détecter une différence de taille moyenne entre les deux populations de 0,01 mm sachant que l'écart-type dans chaque population est de 1 mm ?

2 Taux de fausse découverte et méthode de Benjamini-Hochberg

Le but fondamental du test statistique tel qu'introduit par Fisher est d'exprimer le niveau de preuve expérimentale *contre* une hypothèse nulle, via la p-valeur. Attention : en tant que telle, cette démarche ne répond pas à la question de savoir s'il vaut mieux croire que $\mathcal{H}_0$ est fausse ou non. Pour cela, on comprend qu'il faudrait savoir quelles sont les alternatives, quelles sont les connaissances préalables sur la question posée ou encore quels sont les coûts associés en cas d'erreur. La démarche de test proposée par Fisher n'abordant pas ces aspects, on utilise généralement la p-valeur dans le cadre d'une discussion faisant intervenir (informellement) d'autres éléments. Dans une communication scientifique, on prendra donc bien garde de rapporter dans ses résultats les p-valeurs obtenues pour les différents tests menés, *en précisant bien à chaque fois le nom du test ou l'hypothèse nulle testée*, et de différer l'interprétation du test au niveau de la discussion.

On pourrait résumer les choses en disant qu'il n'existe pas de moyen général pour déterminer un seuil en dessous duquel il vaudrait mieux rejeter l'hypothèse nulle [1]. Une exception notable concerne les situations où, au lieu d'un test isolé, on considère une collection de tests. Ces tests peuvent typiquement venir d'une même expérience [2], lorsque l'on réalise plusieurs mesures en parallèle (comme en transcriptomique par exemple). Le cadre théorique pour traîter ce genre de situation est appelé *test multiple*, et nous allons voir dans la section suivante qu'il permet de dériver un seuil de p-valeur à partir d'un autre critère souhaitable en pratique.

2.1 Fausses découvertes

Pour rappel, lorsqu'on se fixe un seuil de rejet α pour les p-valeurs d'un test, on peut caractériser les réponses à une série de tests de la manière suivante :

	$\mathcal{H}_0$ fausse	$\mathcal{H}_0$ vraie
$P \leq \alpha$	vrai positif	faux positif
$P > \alpha$	faux négatif	vrai négatif

Si au lieu d'un seul test on en a tout un ensemble, spécifier un seuil α revient à filtrer le sous-ensemble des tests qu'on juge significativement en défaveur de l'hypothèse nulle. Dans le cas d'une analyse transcriptomique, il s'agit des gènes différentiellement exprimés, et ce sont les « découvertes » de l'expérience. Dans le lot, il est possible que se trouvent des faux positifs, et on voudrait pouvoir, sinon éviter leur présence, au moins contrôler leur proportion dans la liste des découvertes que l'on publie (i.e. ne pas dépasser une certaine valeur).

Ainsi on définit le **taux de fausse découverte** (*False Discovery Rate, FDR*) pour un seuil de p-valeur par :

$$FDR = \frac{\text{faux positifs}}{\text{positifs}} = \frac{\text{faux positifs}}{\text{vrais positifs + faux positifs}}$$

Contrairement à la p-valeur, il s'agit d'une grandeur aisément interprétable : par exemple toujours dans le cas d'une analyse d'expression différentielle en transcriptomique, elle indique la proportion de gènes sortant significatifs et qui pourtant ne sont pas différentiellement exprimés.

Nous allons voir comment on peut, à l'aide d'un ensemble de tests, déterminer un seuil α sur les p-valeurs qui garantit qu'en moyenne on ne dépasse pas un certain taux de fausse découverte. Mais d'abord un exercice pour mettre en application la notion de fausse découverte.

Exercice 8
Nous proposons une simulation permettant d'apprécier les garanties offertes par un seuil de p-valeur à 0,05.

1. générez 10000 échantillons de taille 10 tirés de $N(0,1)$ et 1000 de $N(1,1)$ et collectez les p-valeurs contre $\mathcal{H}_0 : X \sim N(0,1)$ avec un test t.
2. tracez l'histogramme des p-valeurs obtenues (utilisez l'option `breaks`)
3. on place le seuil de significativité à 0,05. Calculez le taux de fausse découverte
4. calculez le taux de fausse découverte en vous restreignant aux p-valeurs comprises entre 0,04 et 0,05

1. Cela exclut en particulier la « règle » $p < 0.05$ (*cf* exercice 8).
2. On verra qu'il n'y a pas d'obligation formelle à cela.

5. conclusion ?
6. en combinant deux appels à `hist` (avec l'option `add = T`), représentez la proportion d'hypothèses nulles fausses en dans chaque barre de l'histogramme de la question 2.
7. déduisez de la figure générée à la question précédente une méthode géométrique pour estimer le taux de fausse découverte pour un seuil de p-valeur donné

2.2 Méthode de Benjamini-Hochberg

Nous allons maintenant présenter une méthode permettant de contrôler le taux de fausse découverte, appelée **méthode de Benjamini-Hochberg**.

On appelle mélange de deux distributions la distribution obtenue en piochant aléatoirement dans l'une ou l'autre des distributions selon une certaine proportion. Pour simuler le mélange de deux distributions d_1 et d_2 selon une proportion π, on simule $X_1 \sim d_1$, $X_2 \sim d_2$, $B \sim \text{Bern}(\pi)$ et on calcule $X = BX_1 + (1-B)X_2$. Pour obtenir la fonction de répartition F du mélange à partir des fonctions de répartition F_1 et F_2 de d_1 et d_2 et de la proportion π du mélange, on remarque que :

$$\begin{aligned} F(t) &= \mathbb{P}[X \leq t] \\ &= \mathbb{P}[B=1]\mathbb{P}[X \leq t | B=1] + \mathbb{P}[B=0]\mathbb{P}[X \leq t | B=0] \\ &= \mathbb{P}[B=1]\mathbb{P}[X_1 \leq t] + \mathbb{P}[B=0]\mathbb{P}[X_2 \leq t] \\ &= \pi F_1(t) + (1-\pi)F_2(t) \end{aligned}$$

La distribution des p-valeurs pour une collection de tests peut être décrite comme un mélange, avec d'une part les p-valeurs venant des tests où l'hypothèse nulle est vraie, et d'autre part celles venant de tests où l'hypothèse nulle est fausse. En vertu de ce que l'on vient de voir la fonction de répartition F des p-valeurs est donc un mélange d'une distribution uniforme F_0 et d'une distribution F_1 inconnue, telles que :

$$\forall x \in \mathbb{R}\, F(x) = \pi_0 F_0(x) + (1-\pi_0)F_1(x)$$

où π_0 est la proportion de tests où l'hypothèse nulle est vraie dans la collection. Notons bien que $F(x)$ représente la proportion de p-valeurs inférieures à x dans la population totale, et $F_0(x)$ la proportion parmi les valeurs générées sous $\mathcal{H}_0$ de celles inférieures à x. Puisque les p-valeurs générées sous $\mathcal{H}_0$ sont distribuées uniformément, on a $F_0(x) = x$ pour tout x.

Avec ces notations on peut écrire le taux de fausse découverte en fonction d'un seuil x :

$$\text{FDR}(x) = \frac{\pi_0 F_0(x)}{F(x)}$$

d'où l'on tire :

$$\text{FDR}(x) \leq \frac{F_0(x)}{F(x)}$$

puisque $\pi_0 \in [0;1]$. Cette borne est à la base de la méthode de Benjamini-Hochberg **[?]**, qui prend en entrée une collection de p-valeurs, et estime une majoration du FDR pour différents seuils. Nous la décrivons comme suit : soit $P_1, \ldots, P_n$ une collection de p-valeurs issues d'un test

multiple. On les suppose ordonnées de la plus petite à la plus grande. La méthode de Benjamini-Hochberg permet d'estimer une borne supérieure pour FDR(x) où $x \in \{P_1, \ldots, P_n\}$. Elle utilise pour cela l'estimateur suivant :

$$F(P_i) = \frac{i}{n}$$

et comme par ailleurs on a $F_0(P_i) = P_i$, on obtient comme borne de FDR(P_i) la valeur $p_i \frac{n}{i}$. La méthode de Benjamini-Hochberg comporte une dernière étape :

$$Q_i = \begin{cases} P_i & \text{si } i = n \\ \min\{Q_{i+1}, P_i \frac{n}{i}\} & \text{sinon} \end{cases}$$

La valeur Q_i a ainsi pour propriété de majorer le taux de fausse découverte $FDR(P_i)$. Cette estimation du taux de fausse découverte est excessivement utile en pratique. En effet s'il est à peu près impossible de donner un seuil de p-valeur s'adaptant à tous les tests/tailles d'échantillonnage, il est en revanche très intuitif de fixer un taux de fausse découverte maximum tolérable en fonction du contexte. Par exemple dans une analyse d'expression différentielle, on détecte typiquement plusieurs centaines voire quelques milliers de gènes. On peut tolérer entre 1 et 10% de fausse découverte en fonction du contexte plus ou moins exploratoire et de l'utilisation faite en aval de la liste trouvée.

Exercice 9
Calculez les estimations de taux de fausse découverte selon la méthode de Benjamini-Hochberg pour les p-valeurs suivantes : 0,91 0,11 0,71 0,31 0,51 0,41 0,61 0,21 0,81 0,01. Vérifiez vos résultats à l'aide de la fonction `p.adjust`.

La méthode de Benjamini-Hochberg est dite « conservative », parce qu'elle garantit qu'en moyenne l'estimation du taux de fausse découverte est supérieure à la vraie valeur. Il s'agit donc d'une estimation pessimiste en moyenne. Ceci est d'ailleurs illustré dans l'exercice suivant.

Exercice 10
On cherche ici à évaluer sur des simulations l'écart entre l'estimation de FDR par la méthode de Benjamini-Hochberg et le vrai taux de fausse découverte.

1. en reprenant la simulation de l'exercice 8, calculez la valeur empirique du taux de fausse découverte pour chaque p-valeur générée dans la simulation. Notez que cela est possible parce qu'ayant réalisé la simulation, vous savez pour chaque p-valeur si elle a été générée sous l'hypothèse nulle ou non.
2. Calculez les estimations de FDR par la méthode de Benjamini-Hochberg (utilisez la fonction `p.adjust`)
3. générez un graphique superposant l'estimation obtenue avec le vrai taux de fausse découverte
4. commentaires ?

2.3 Un point de vocabulaire

Il est à noter que la méthode de Benjamini-Hochberg est très souvent désignée avec d'autres méthodes (comme celle de Bonferroni) par le terme d'ajustement ou de correction de p-valeur.

L'auteur de ces lignes réprouve fortement l'usage de ces termes, qui sont utilisés avec l'idée que les p-valeurs sont « faussées » par le test multiple. Comme on l'a vu, il n'en est rien : le test multiple ne « fausse » rien, mais permet de calculer, à partir de la distibution empirique des p-valeurs sur un grand nombre de cas, une grandeur beaucoup plus commode que la p-valeur, le taux de fausse découverte.

3 Méthode de Bonferroni

Avec la méthode de Benjamini-Hochberg, on a un moyen d'exploiter la situation de test multiple pour contrôler la proportion de faux positifs dans une liste de résutats. Selon le contexte, on peut avoir besoin d'une propriété plus forte, à savoir contrôler pour la présence ne serait-ce que d'un seul faux-positif.

Pour les exercices et explications qui suivent, on se donne le paysage suivant. Soit une collection d'hypothèses nulles $H_1, \ldots, H_m$. Insistons bien : on les appelle $H_1, \ldots, H_m$ mais ce sont toutes des hypothèses nulles, chacune associée à son test. On produit ainsi une collection $P_1, \ldots, P_m$ de p-valeurs.

Exercice 11
On suppose ici que toutes les hypothèses nulles sont vraies, et on se donne un seuil α en dessous duquel on décide de rejecter l'hypothèse nulle.

1. Comment dénote-t-on la probabilité d'avoir un faux-positif sur le premier test sachant que l'hypothèse nulle correspondante est vraie ? Combien vaut cette probabilité ?
2. Quelle est la probabilité de ne pas avoir de faux-positifs sur le premier test sachant que l'hypothèse nulle correspondante est vraie ?
3. Quelle est la probabilité d'avoir 0 faux-positifs sur les 2 premiers tests sachant que les hypothèses nulles sont vraies ?
4. Quelle est la probabilité d'avoir 0 faux-positifs quand on réalise les m tests sachant que toutes les hypothèses nulles sont vraies ?
5. Déterminez la limite pour $m \to +\infty$ de l'expression trouvée à la question précédente. Proposez une représentation graphique et commentez.

Dans la suite de cette section, on note $m_0 \leq m$ le nombre d'hypothèses nulles vraies parmi $H_1, \ldots, H_m$, et on peut sans perte de généralité supposer que ce sont les m_0 premières.

Posons-nous maintenant la question suivante : y a-t-il un moyen de choisir un seuil de significativité pour les tests de manière à garantir qu'en moyenne la probabilité qu'il y ait au moins un faux positif parmi les m soit au plus α ? La **méthode Bonferroni** est une approche très simple pour parvenir à ce résultat. Elle consiste à choisir une seuil de significativité de $\frac{\alpha}{m}$ pour les tests. Commençons par la tester expérimentalement.

Exercice 12
Déterminez un protocole expérimental pour mettre à l'épreuve la méthode Bonferroni, et mettez-le en œuvre.

Exercice 13
Nous allons ici démontrer que la méthode de Bonferroni permet bien de contrôler la probabilité d'un faux positif parmi m tests.

1. Donnez une expression du taux de faux positifs en fonction des P_i et de α.
2. Donnez une expression du taux de fausse découverte en fonction des P_i et de α. Comparez à l'expression précédente et proposez un schéma pour illustrer la différence entre les deux notions.
3. Écrivez l'évènement correspondant à $A \equiv$ « au moins un faux positif parmi les m tests pour un seuil $\frac{\alpha}{m}$ ».
4. Qu'est-ce que l'inégalité de Boole ? De quelle règle plus fondamentale de calcul des probabilité peut-on la faire dériver ?
5. Utilisez l'inégalité de Boole pour démontrer que la probabilité de l'évènement A est inférieure à α.

Bonferroni ou Benjamini-Hochberg ? À ce stade, nous avons vu deux procédures permettant d'exploiter une collection de valeurs p. Celles-ci sont souvent désignées comme des méthodes de « correction pour le test multiple », mais insistons bien encore une fois qu'il n'y a ici aucune « correction ». Il faut bien comprendre qu'entre un test isolé, une application de Benjamini-Hochberg sur une collection de p-valeurs ou de celle de Bonferroni, on pose à chaque fois une question différente et qu'il n'y a pas lieu de les mettre sur le même plan :

— pour un test isolé, on demande quelle est la probabilité d'observer un résultat plus extrême que celui observé sous l'hypothèse nulle

— dans la méthode de Benjamini-Hochberg, on recherche un seuil de significativité pour les tests individuels d'une collection qui contrôle le taux de fausse découverte

— dans la méthode de Benjamini, on recherche un seuil de significativité pour les tests individuels d'une collection qui contrôle la probabilité de présence ne serait-ce que d'un seul faux positif.

4 Méthode de Holm-Bonferroni

On s'intéresse ici à une autre méthode permettant de contrôler la probabilité d'existence d'un faux positif, appelée **méthode de Holm-Bonferroni**. En voici une description.

On suppose maintenant les p-valeurs $P_1, \ldots, P_m$ ordonnées et on procède itérativement pour déterminer les tests dont on rejette l'hypothèse nulle :

— si $P_1 < \frac{\alpha}{m}$ rejeter H_1 sinon terminer

— si $P_2 < \frac{\alpha}{m-1}$ rejeter H_2 sinon terminer

— ...

— si $P_i < \frac{\alpha}{m+1-i}$ rejeter H_k sinon terminer

On peut prouver qu'en procédant ainsi on garantit que la probabilité d'avoir au moins un faux positif parmi les m tests est inférieure à α et qu'en plus cette procédure rejette plus d'hypothèses que celle de Bonferroni.

Exercice 14
Mettez ces deux affirmations à l'épreuve par le protocole expérimental de votre choix.

Exercice 15
Démontrez que la méthode de Holm-Bonferroni est effectivement plus puissante que celle de Bonferroni (elle rejette plus d'hypothèses nulle pour un même niveau de contrôle).