Plan

Exemple introductif

- On imagine une urne contenant un nombre infini de balles.
- Les balles sont soit blanches, soit rouges.
- On note p la proportion de balles rouges.

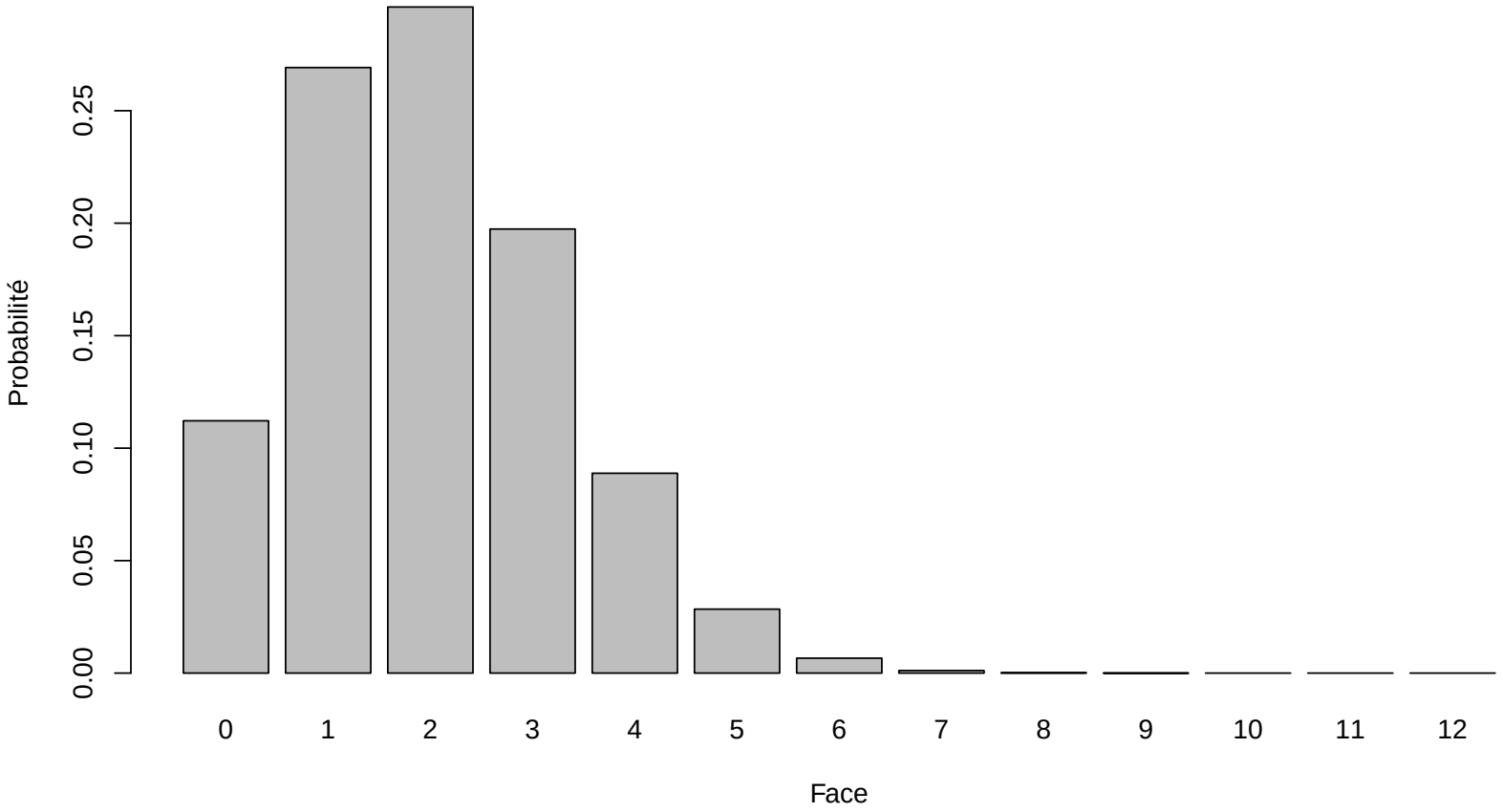
- Considérons l'expérience consistant à tirer n balles dans l'urne, et rapporter le nombre N de balles rouges obtenues.
- En moyenne, on doit obtenir np balles rouges.
- Mais on veut être plus précis et calculer :
  - la probabilité d'obtenir 0 balle rouge sur n tirages,
  - la probabilité d'obtenir 1 balle rouge sur n tirages,
  - la probabilité d'obtenir 2 balles rouges sur n tirages,
  - ...

Un peu de vocabulaire

- le tirage des n balles constitue une expérience aléatoire
- dont le résultat est le nombre N de fois où on a obtenu une balle rouge
- N est appelée variable aléatoire
- la moyenne de N est la valeur type à laquelle s'attendre lorsqu'on réalise l'expérience
- la distribution de probabilité de N est la probabilité que N vaille k pour $k = 0, 1, \ldots, n$

Loi binomiale

La variable aléatoire N suit une loi binomiale :



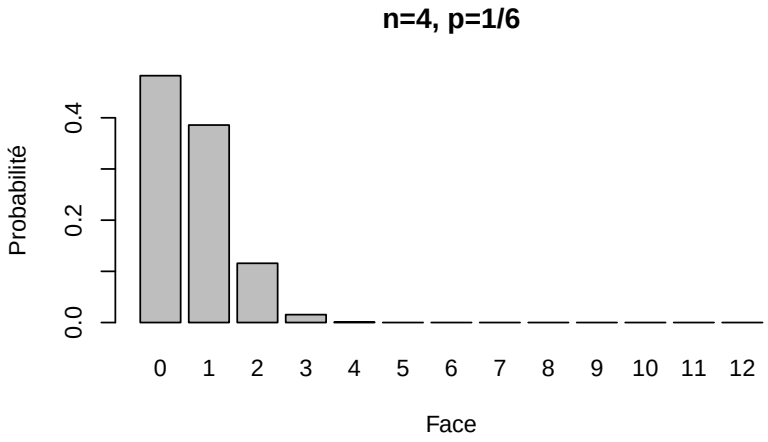
Définition un peu plus abstraite

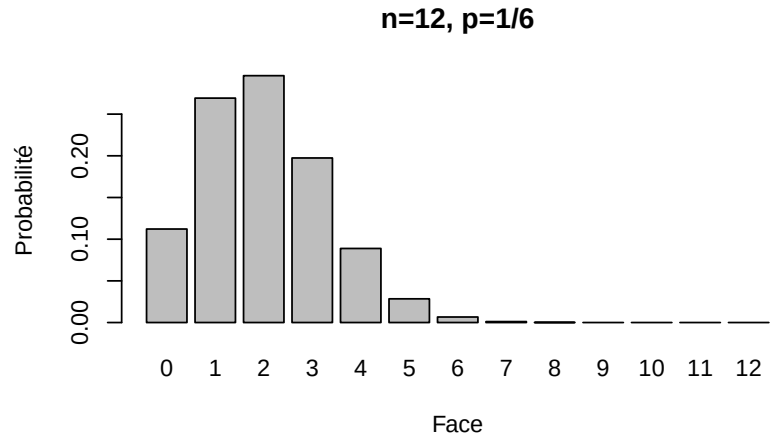
- On considère une épreuve E qui peut réussir ou échouer
- de manière indépendante d'une fois sur l'autre
- si la probabilité de succès est p
- et le nombre d'essais est n
- alors N le nombre de succès sur n essais suit une loi binomiale de paramètres n et p.
- pour faire savant on écrit :

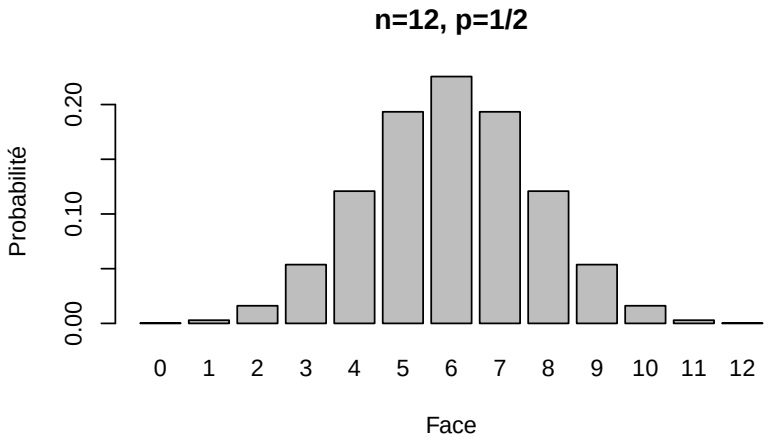
$$N \sim B(n, p)$$

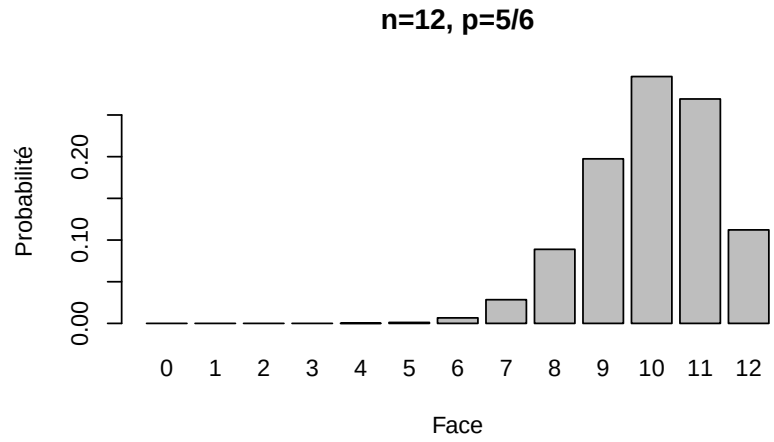
La famille des lois binomiales

En changeant les paramètres n et p on obtient une famille de distributions :





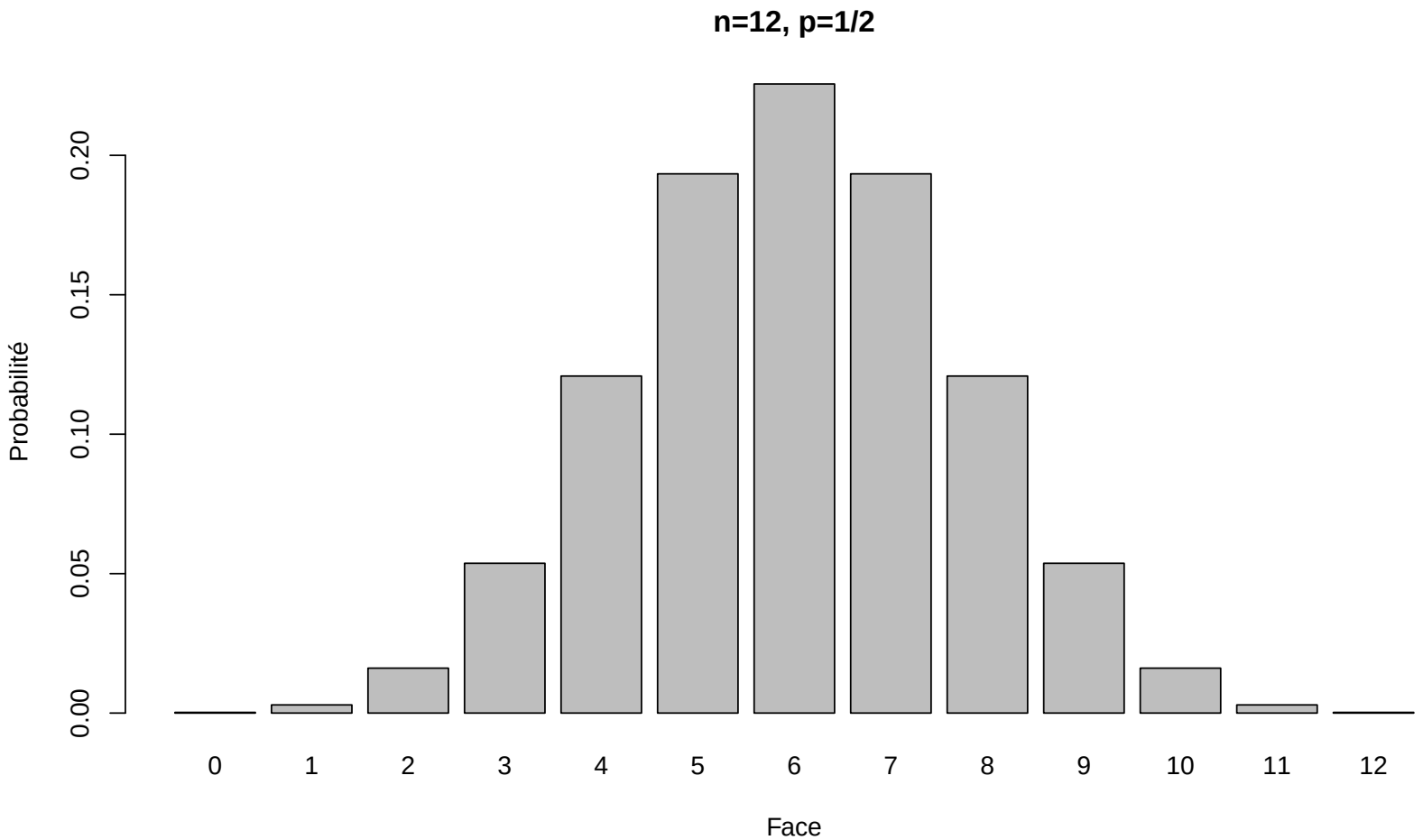




Bon à savoir

Pour $N \sim B(n, p)$,

- Moyenne : $\mathbb{E}[N] = np$
- Variance : $\mathbb{V}[N] = np(1-p)$



Et les NGS dans tout ça ?

balle	**lecture**
tirage dans l'urne	**séquençage + alignement**
nombre de tirages n	**profondeur de séquençage**
nombre de balles rouges sur n tirages	**nombre de lectures dans une région d'intérêt**
prob. des balles rouges p	**proportion de lectures dans une région d'intérêt**

Que mesure-t-on avec les NGS ? (1/2)

- Soit une région génomique (p. ex. un gène, un promoteur ...)
- On pose N le nombre de lectures tombant dans cette région
- $N \sim B(n, p)$
- On s'attend en moyenne à np lectures

- Normalisation ?
  - par la profondeur de séquençage $\Rightarrow$ division par $n/10^6$
  - par la taille l de la région $\Rightarrow$ division par $l/10^3$

  $\Rightarrow$ unité **RPKM** (*Reads per kilo-base per million*)
  - le signal en RPKM $\frac{10^9 N}{nl}$ vaut en moyenne $\frac{10^9 p}{l}$

Que mesure-t-on avec les NGS ? (2/2)

- Le signal en RPKM est fréquemment (mais incorrectement) assimilé à une mesure de la concentration absolue en milieu cellulaire
- Or ce signal vaut en moyenne p (à une constante près)
- p est la probabilité de tirer une localisation donnée dans la "soupe" des fragments de l'échantillon

> C'est donc une mesure **relative** de l'abondance d'une région donnée.

En particulier : si une région A est constante entre deux conditions mais une espèce B initialement très présente disparait complètement, le niveau **mesuré** de A augmentera.

Inteprétation de p

p reflète l'abondance d'une région dans l'échantillon, modulée par :

- sa susceptibilité à la fragmentation
- l'efficacité de son amplification par PCR
- sa mappabilité sur un génome de référence

Ces aspects sont déterminés par la séquence, mais aussi par l'état chromatinien.

Synthèse 1

Dans une mesure NGS,

- un échantillon est constitué de millions de fragments correspondant à différentes régions d'un génome
- on compte les lectures de fragments provenant d'une région donnée du génome (gène, promoteur, *enhancer*)
- ce nombre une fois normalisé (en RPKM) indique l'abondance relative de la région dans l'échantillon ($\pm$ d'autres effets)
- à état biologique constant, la mesure varie aléatoirement selon une loi binomiale
- ce bruit est intrinsèque au processus d'échantillonage qui a lieu lors du séquençage
- il induit une erreur relative de l'ordre de $\frac{1}{\sqrt{np}}$

Loi de Poisson

Définition intuitive

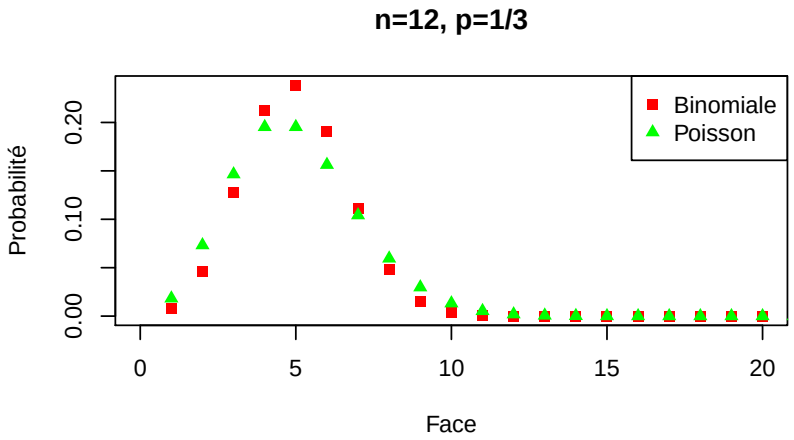
- Soit un évènement E survenant de manière aléatoire et indépendante.
- Supposons que dans un laps de temps donné, E survient *en moyenne* λ fois
- Alors N, le nombre d'occurrences de E dans ce laps de temps suit une loi de Poisson

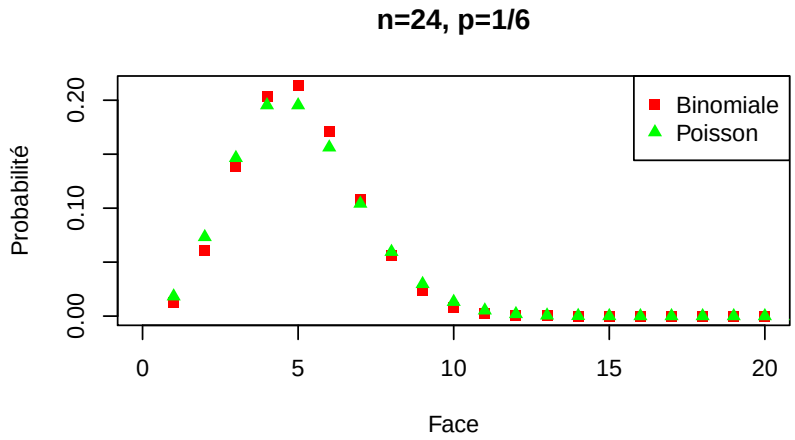
$$N \sim P(\lambda)$$

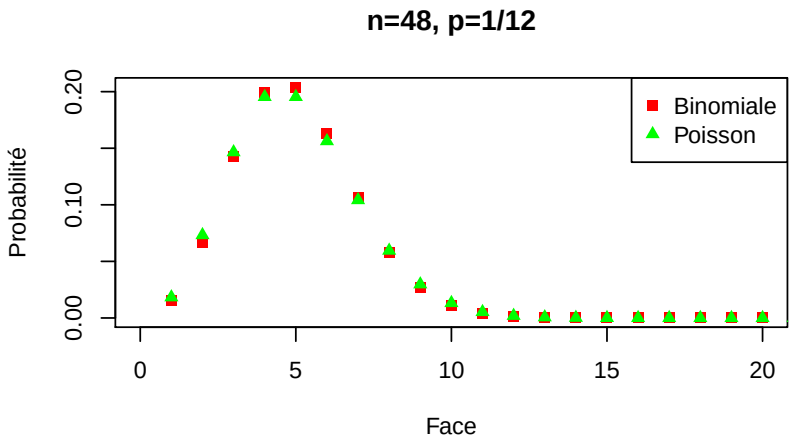
Dans notre contexte :

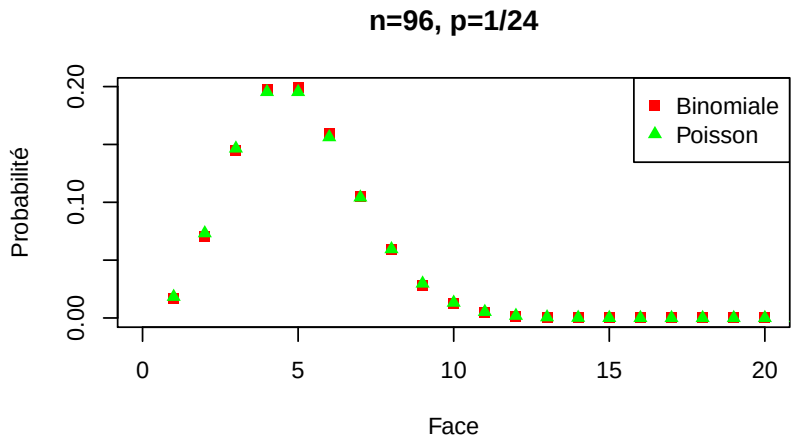
- E = lecture dans une région d'intérêt
- λ = nombre moyen de lectures dans la région pour n lectures séquencées
- λ correspond au produit np des paramètres de la loi binomiale

Convergence de la loi binomiale vers la loi de Poisson









Remarques sur la loi de Poisson

- Utilisée en pratique pour approximer la loi binomiale
- Préférée pour faciliter les calculs
- Si $N \sim P(\lambda)$,

$$\mathbb{E}[N] = \lambda$$
$$\mathbb{V}[N] = \lambda$$

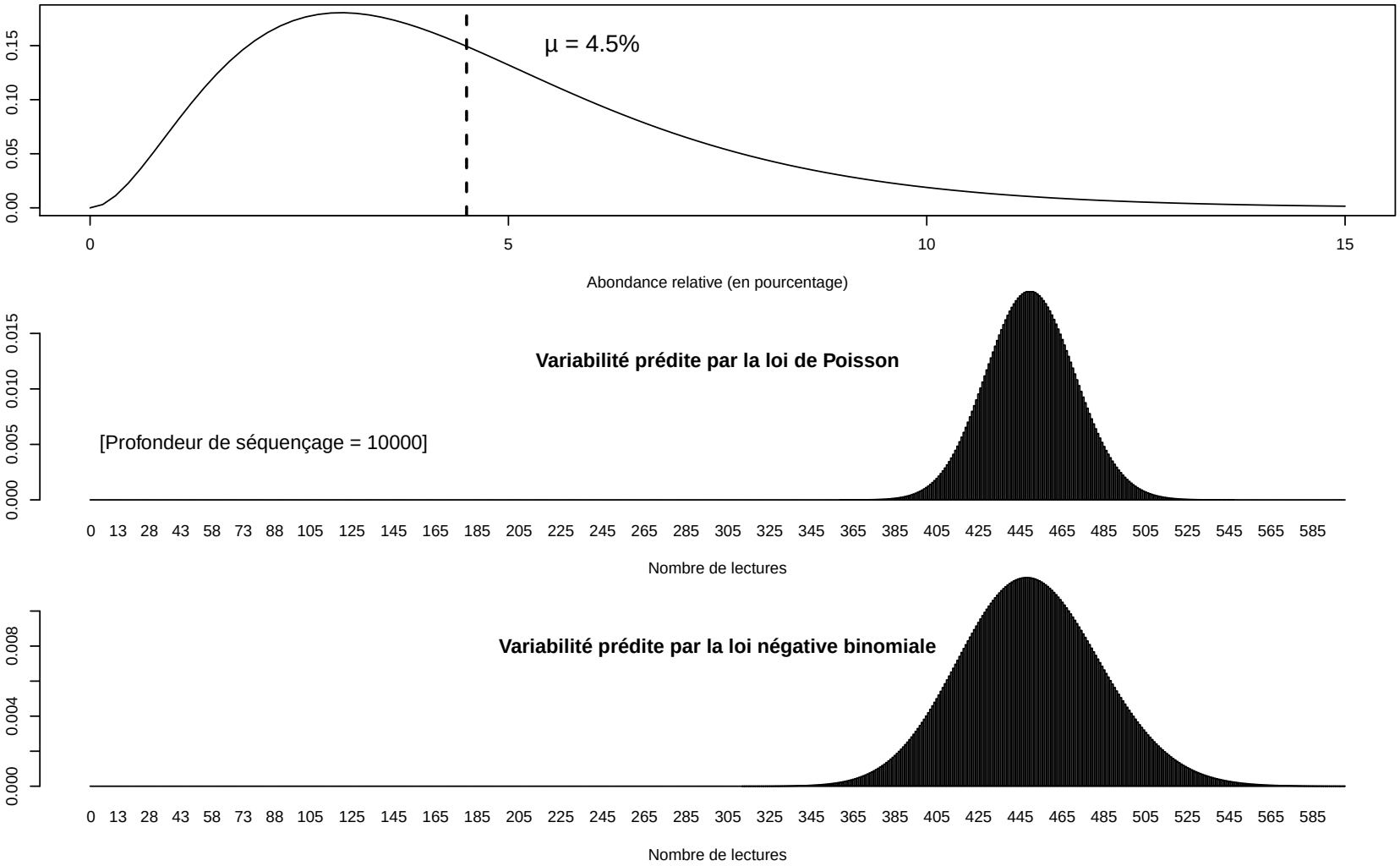
- de même que pour la loi binomiale, la loi de Poisson modélise la variabilité liée à l'échantillonnage

Loi binomiale négative

- en pratique la loi de Poisson décrit bien la variabilité des “réplicats techniques”
  - = séquençages multiples d’un même échantillon
  - $\Rightarrow$ le séquençage induit un bruit négligeable
    - (même s’il a un effet sur p)

- mais sous-estime la variabilité de “réplicats biologiques”
  - = différents échantillons d’une même condition biologique
    - cette variabilité supplémentaire doit être estimée par une méthode statistique à l’aide de réplicats
    - l’estimation sans réplicats est possible mais très fragile (la détection est de variation est beaucoup moins sensible)

Composition et variabilité



Synthèse 2

- La variabilité des mesures NGS provient :
  - d'un bruit "technique" (*shot noise*) lié au séquençage
  - d'un bruit "biologique" (*sample noise*) lié à la préparation de l'échantillon
- La variabilité liée au séquençage :
  - n'est pas un défaut, elle est intrinsèque au processus d'échantillonnage des fragments
  - est fixe à une profondeur de séquençage donnée
  - ne peut être réduite qu'en augmentant la profondeur
- La variabilité liée aux échantillons
  - est spécifique du système étudié
  - ne peut être estimée de manière fiable qu'avec un nombre suffisant de réplicats (> 30 ☺)

Compromis à trouver entre profondeur de séquençage et nombre de réplicats

Plan

Exemple (1/3)

- Soit une région pour laquelle on mesure un signal de ChIP dans deux conditions (contrôle/traitement).
- Le signal est présenté sous la forme d'un nombre de lectures
- On fera abstraction de la différence de profondeur entre librairies en les supposant toujours égales
- Parmi les situations suivantes, dans quel cas a-t-on une "vraie" différence d'expression entre le contrôle et le traitement ?

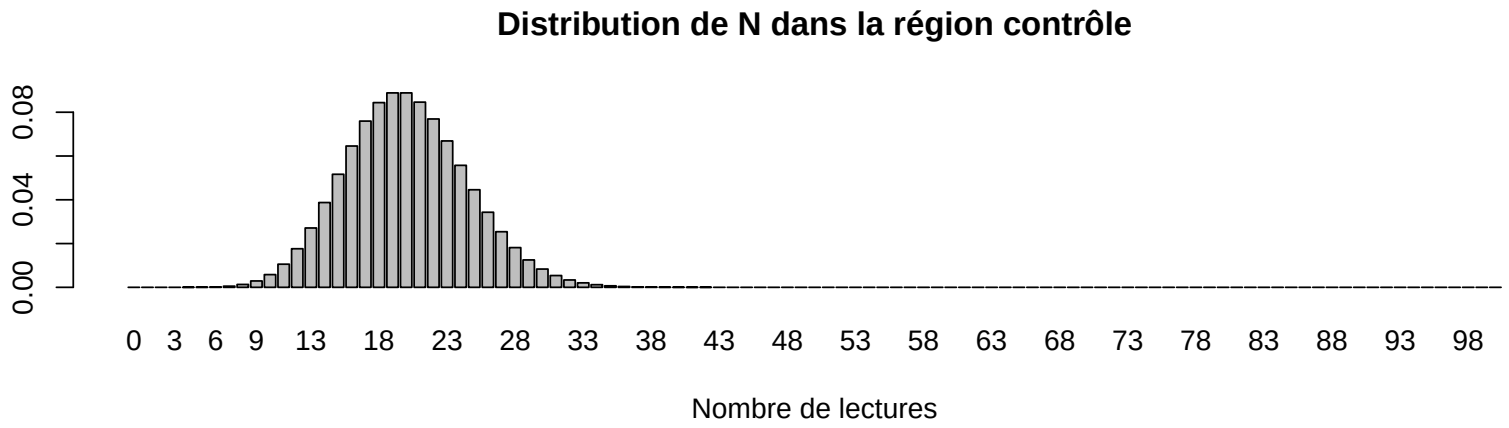
	Contrôle	**Traitement**
S1	3	350
S2	4	5
S3	4	10
S4	12	25
S5	100	140

Exemple (2/3)

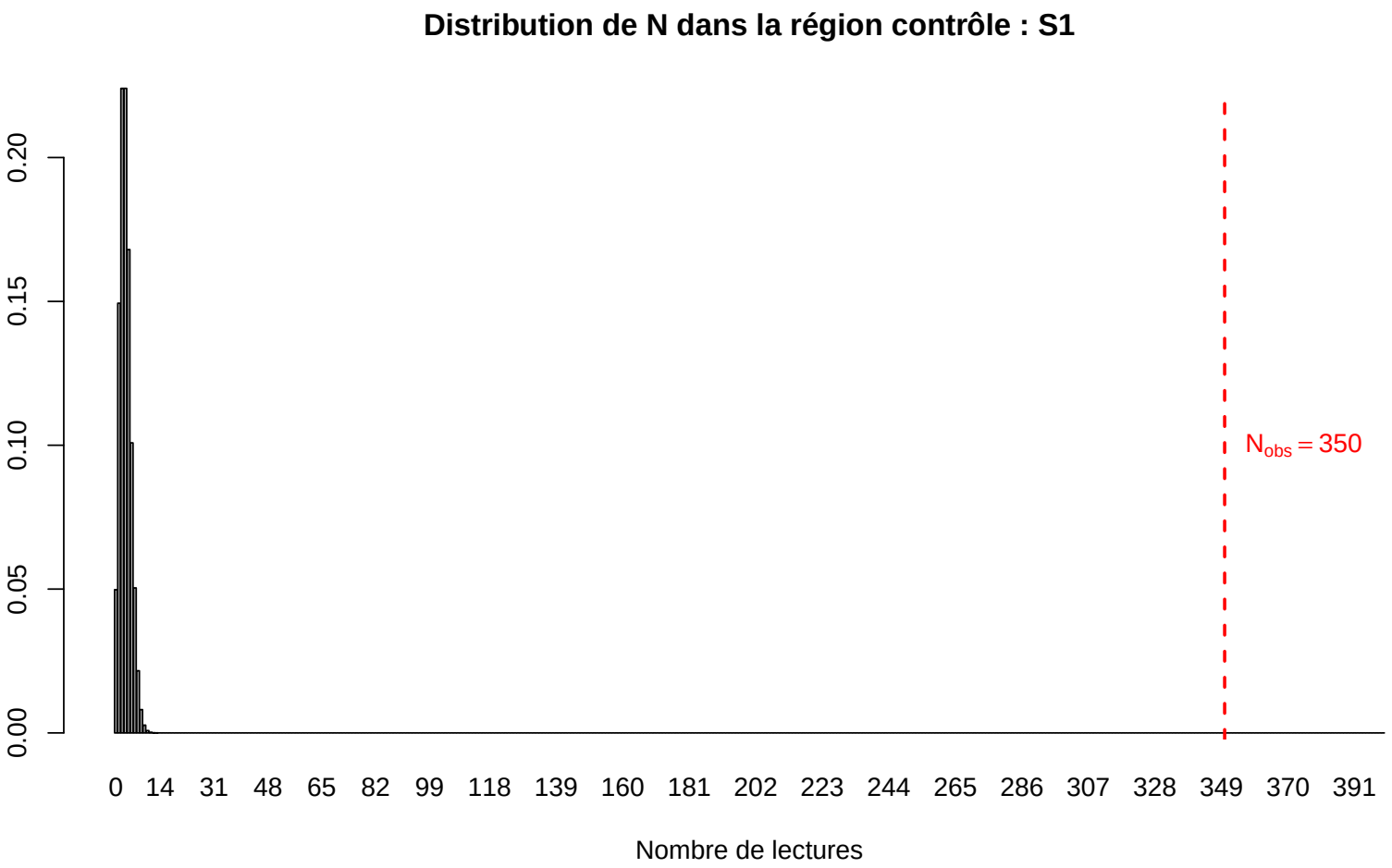
- Il faut tenir compte du *ratio*, mais aussi du nombre absolu de lectures ...
- Pour être plus rigoureux,

  on ne peut pas conclure sans avoir caractérisé la distribution du nombre de lectures dans au moins une condition

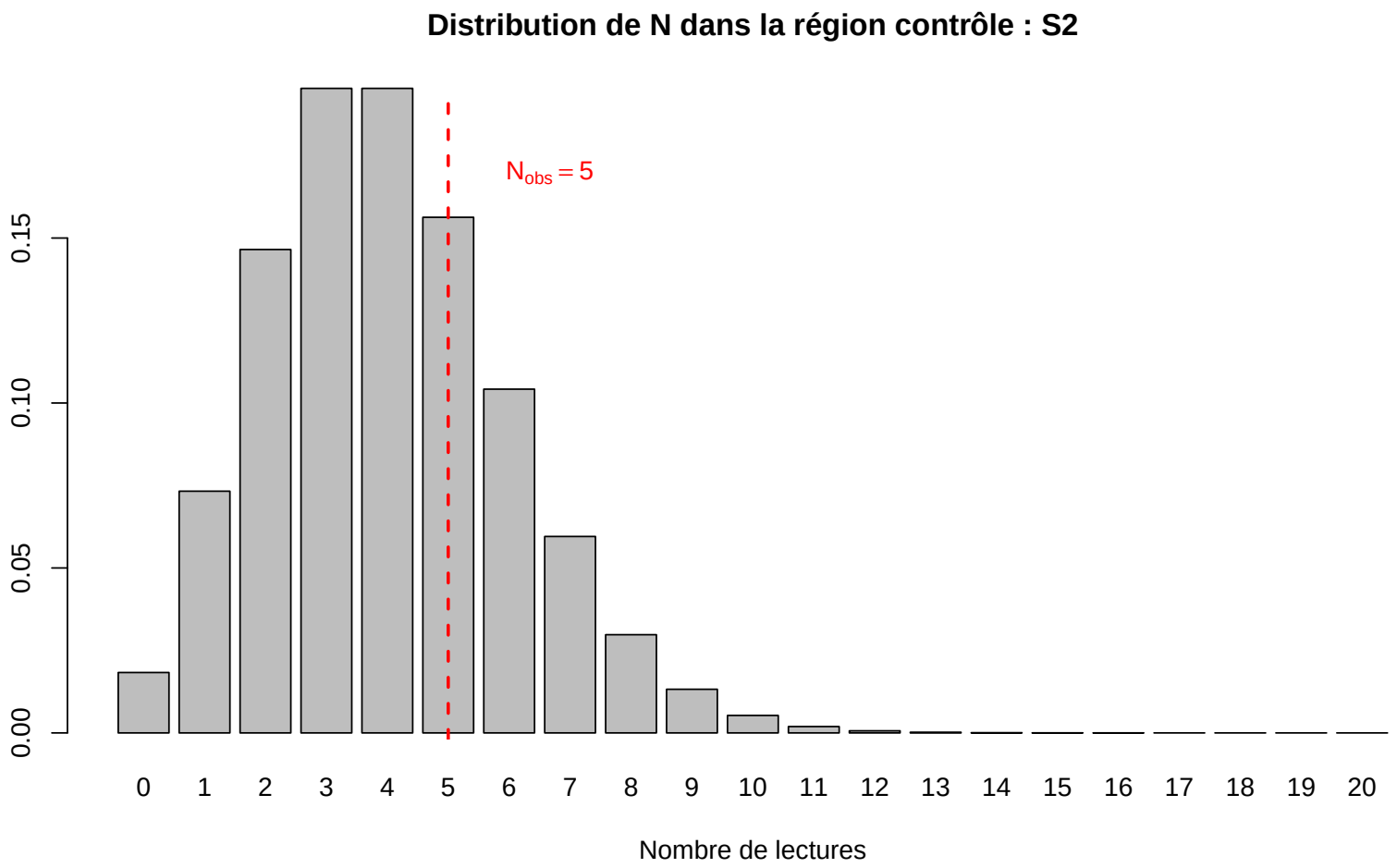
- c'est-à-dire déterminer :



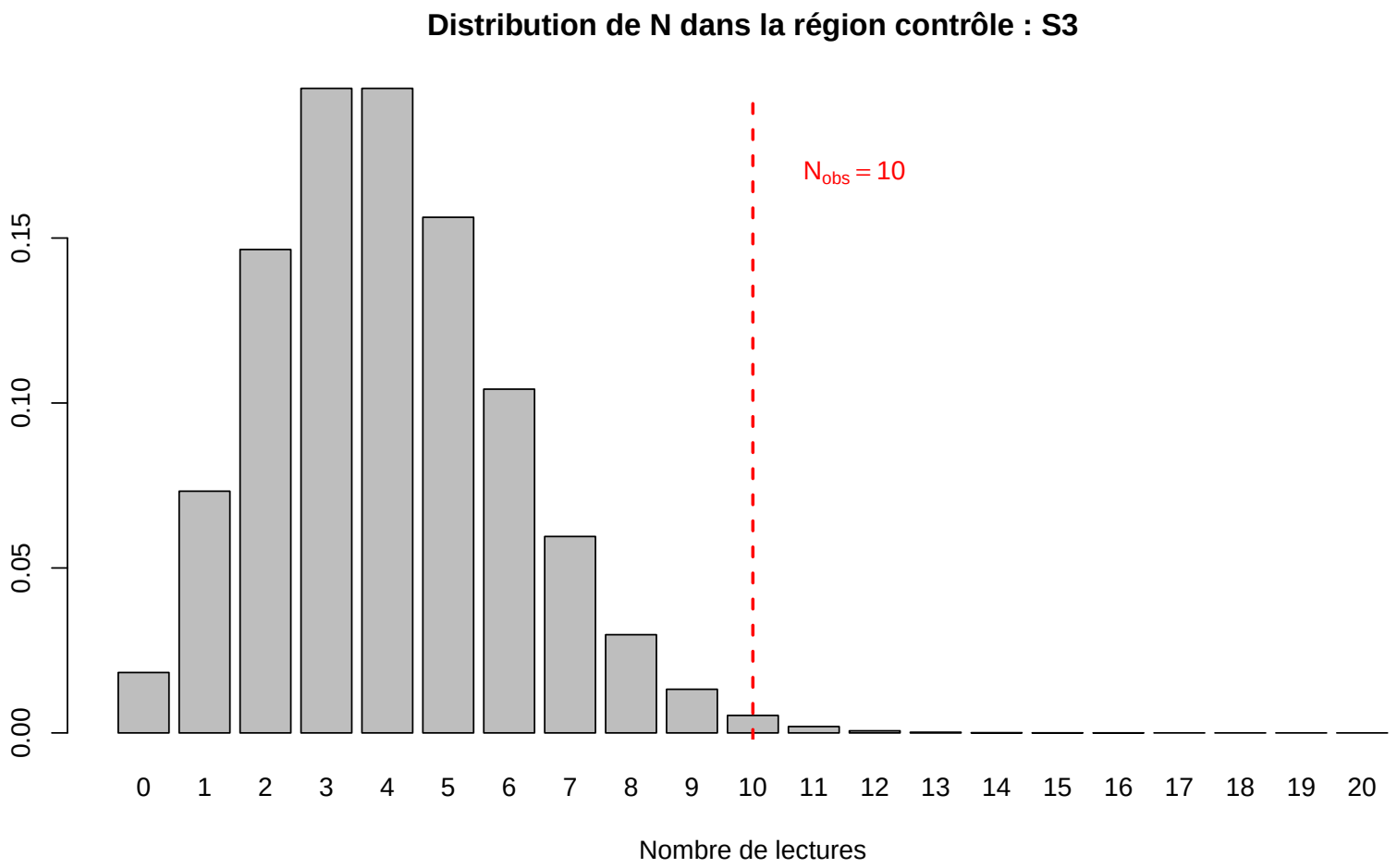
Exemple (3/3)


Distribution de N dans la région contrôle : S1
0.00
0.05
0.10
0.15
0.20
$N_{obs} = 350$
0 14 31 48 65 82 99 118 139 160 181 202 223 244 265 286 307 328 349 370 391
Nombre de lectures

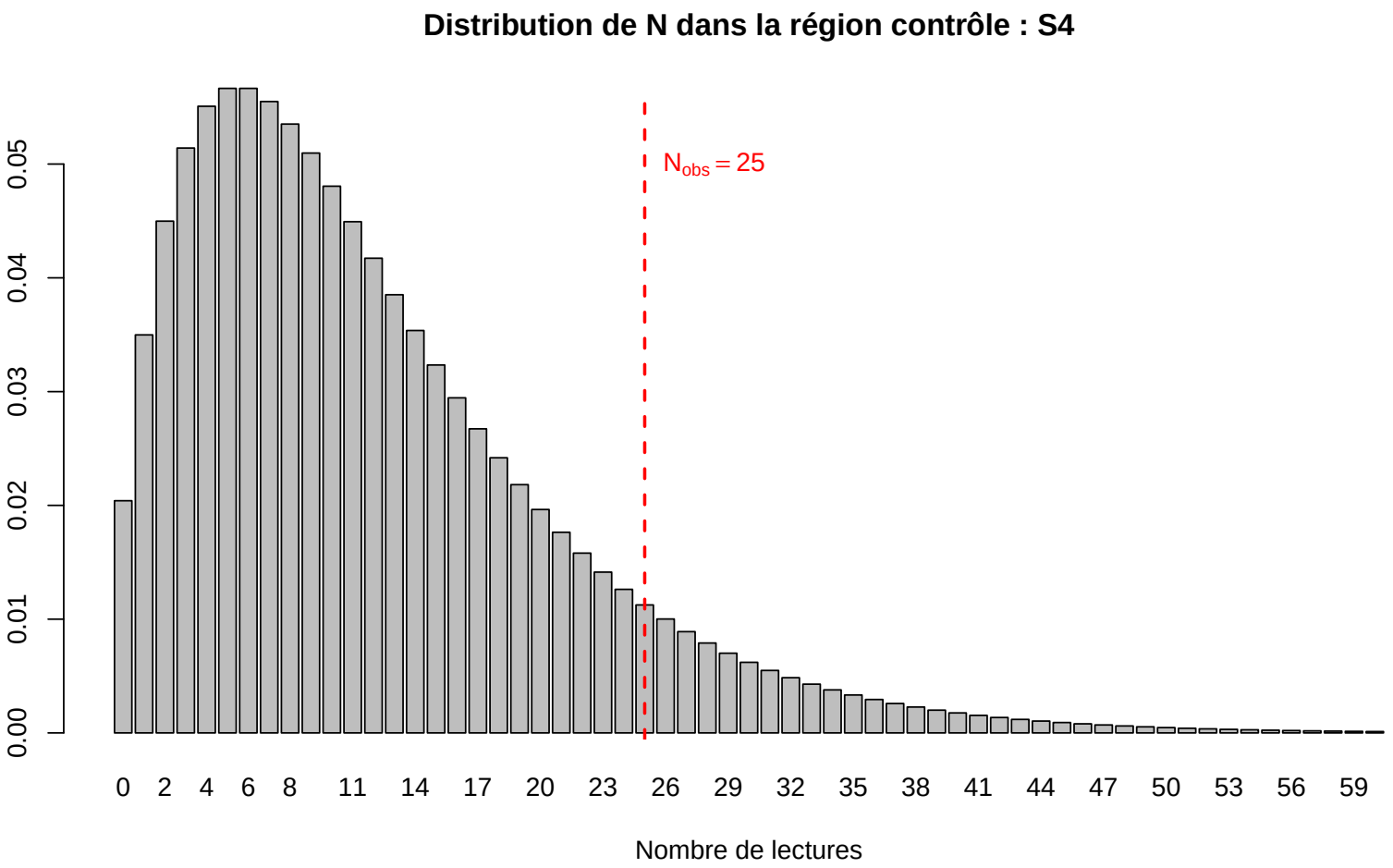
Exemple (3/3)


Distribution de N dans la région contrôle : S2
$N_{obs} = 5$
0.00
0.05
0.10
0.15
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Nombre de lectures

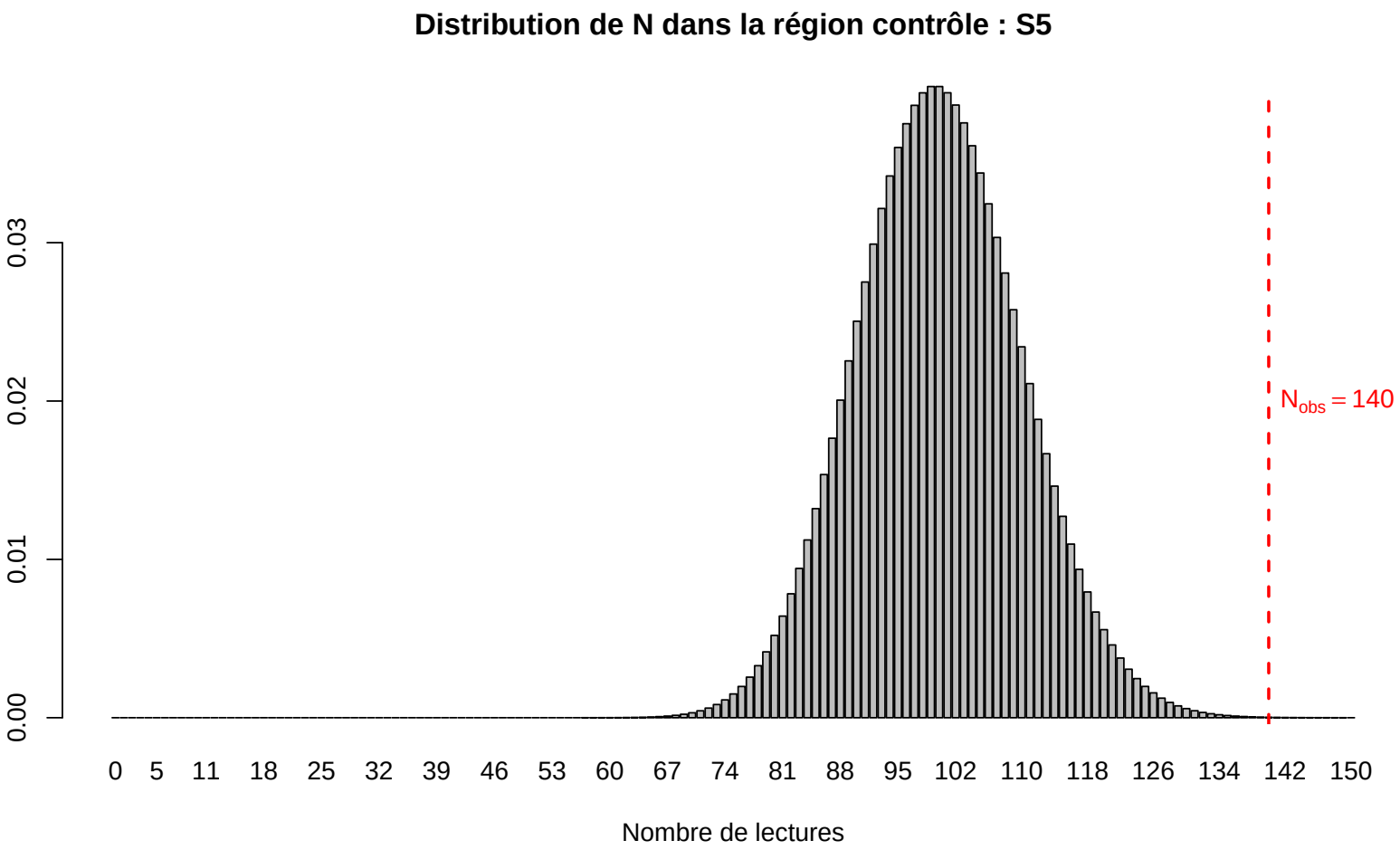
Exemple (3/3)


Distribution de N dans la région contrôle : S3
$N_{obs} = 10$
0.00
0.05
0.10
0.15
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Nombre de lectures

Exemple (3/3)


Distribution de N dans la région contrôle : S4
$N_{obs} = 25$
0.00
0.01
0.02
0.03
0.04
0.05
0 2 4 6 8 11 14 17 20 23 26 29 32 35 38 41 44 47 50 53 56 59
Nombre de lectures

Exemple (3/3)


Distribution de N dans la région contrôle : S5
0.00
0.01
0.02
0.03
$N_{obs} = 140$
0 5 11 18 25 32 39 46 53 60 67 74 81 88 95 102 110 118 126 134 142 150
Nombre de lectures

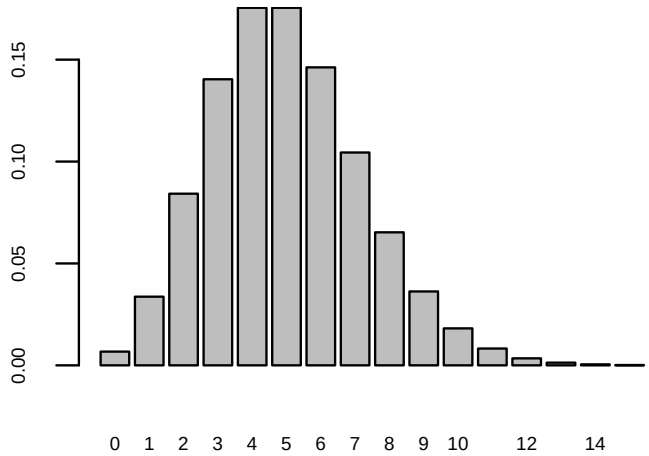
État des lieux : comment interpréter les données ?

- le *ratio* reste un choix pertinent :
  - signification simple à comprendre et à transmettre
  - facile à calculer (que fait-on avec un comptage à zéro dans le contrôle ?)
- mais parfois trompeur
  - comptages faibles
  - et/ou grande variance
- dans l'idéal il faudrait :
  - continuer de raisonner sur les ratios
  - en disposant d'un filtre détectant les cas "non significatifs"

Les tests statistiques jouent ce rôle de filtre

Probabilité des “grandes” valeurs

Dans la condition contrôle, quelle est la probabilité d'avoir $N \geq 10$?



$$\begin{aligned}\mathbb{P}[N \geq 10] &= \mathbb{P}[N = 10] + \mathbb{P}[N = 11] + \mathbb{P}[N = 12] + \dots \\ &= 0.03182806\end{aligned}$$

☞ Plus d'une fois sur 30, on trouvera dans le contrôle une valeur deux fois supérieure à la moyenne

Une démarche de test statistique

Objectif : déterminer une différence d'abondance entre conditions contrôle et traitement

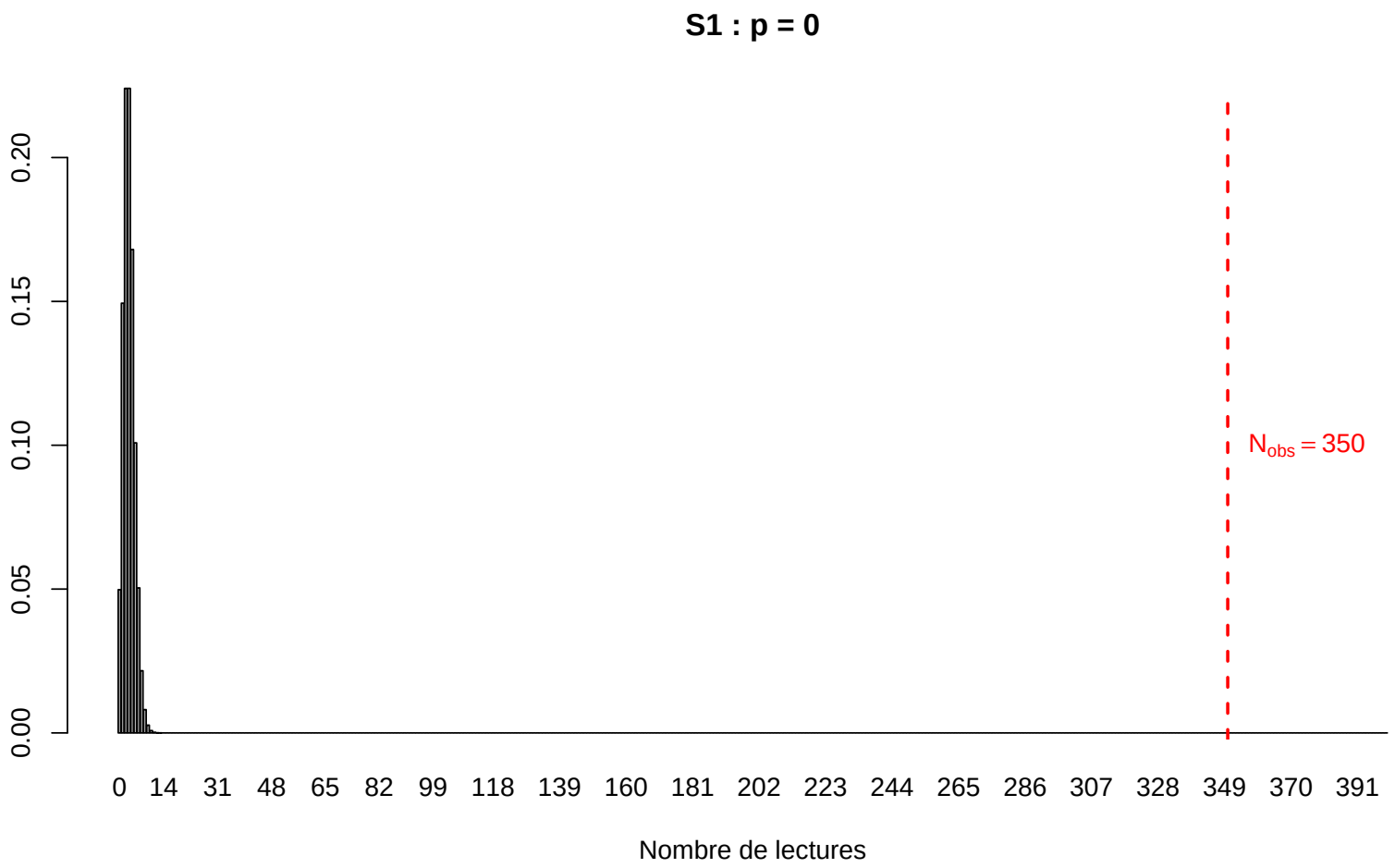
- Caractériser la variabilité dans le contrôle
  = déterminer la distribution de N dans le contrôle
- Mesurer N dans le traitement : n_{obs}
- Supposer que la région étudiée a le même comportement dans le contrôle et le traitement (**hypothèse H_0**)
- Calculer la probabilité (sous H_0) d'obtenir une mesure au moins aussi grande (**p-valeur**)
  - si la p-valeur est grande,

    la variation observée peut très bien s'expliquer par la variabilité attendue dans la condition contrôle

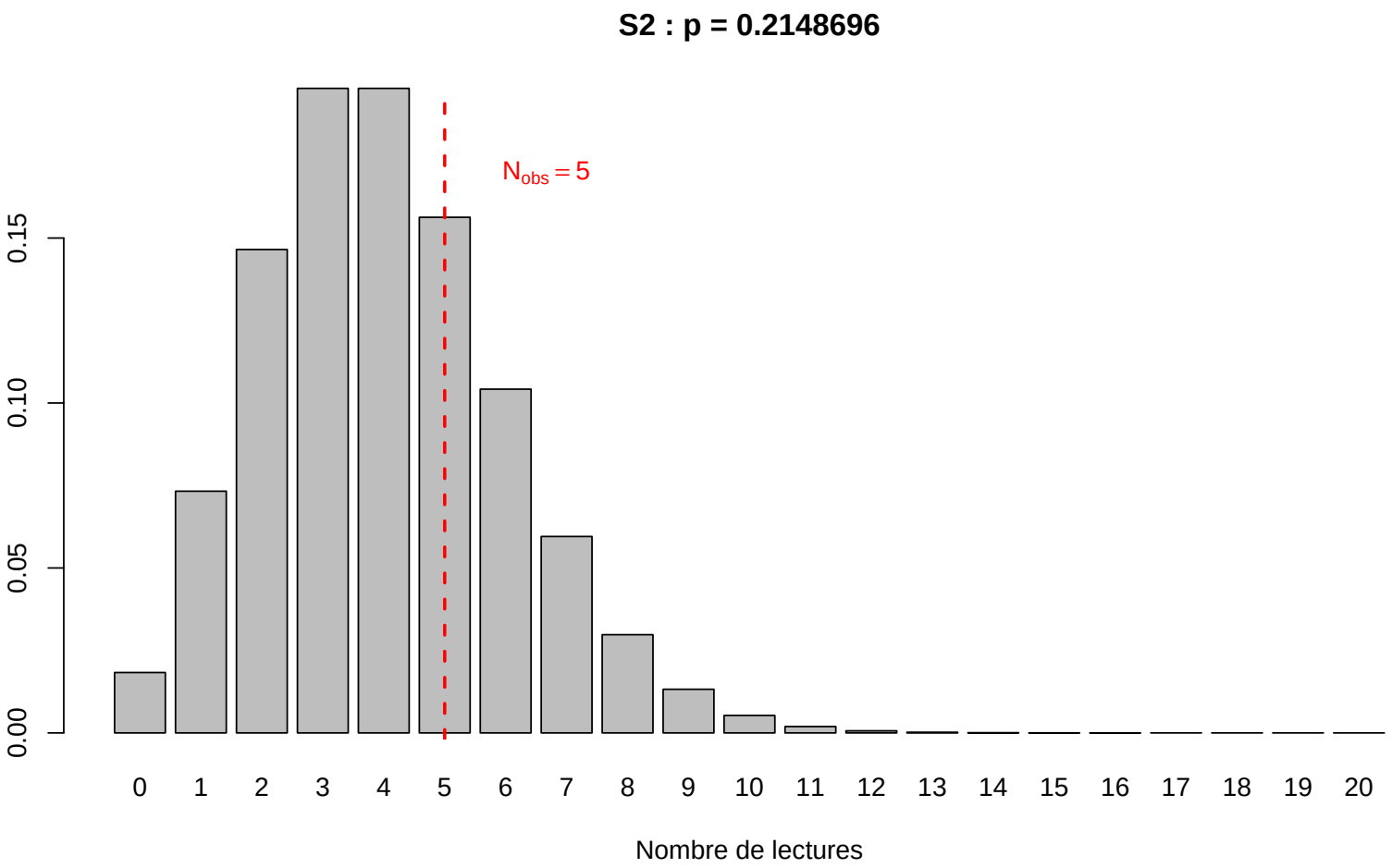
  - si elle très petite

    soit il s'est passé un évènement extrêmement rare, soit l'hypothèse H_0 est **fausse**

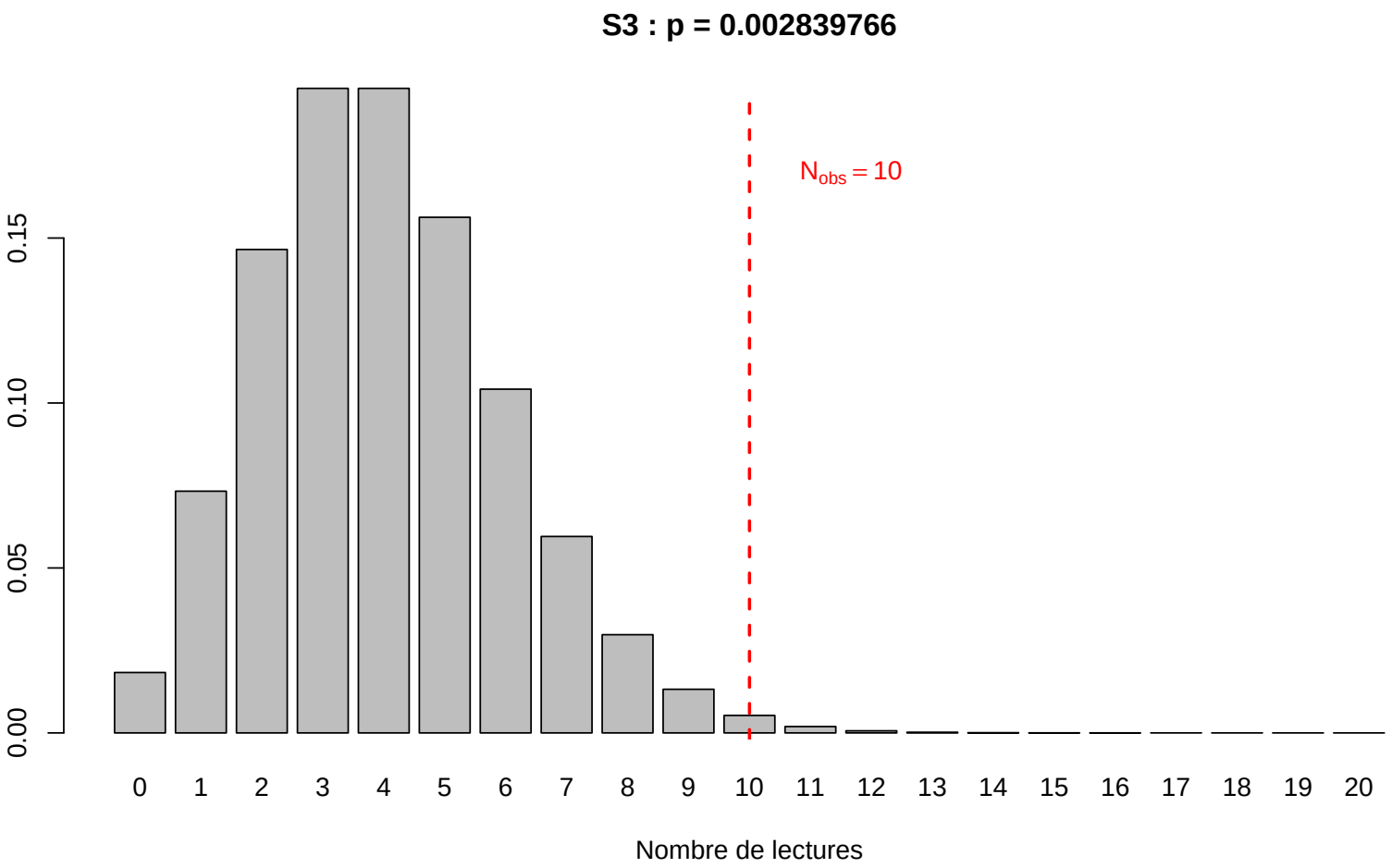
Sur l'exemple précédent



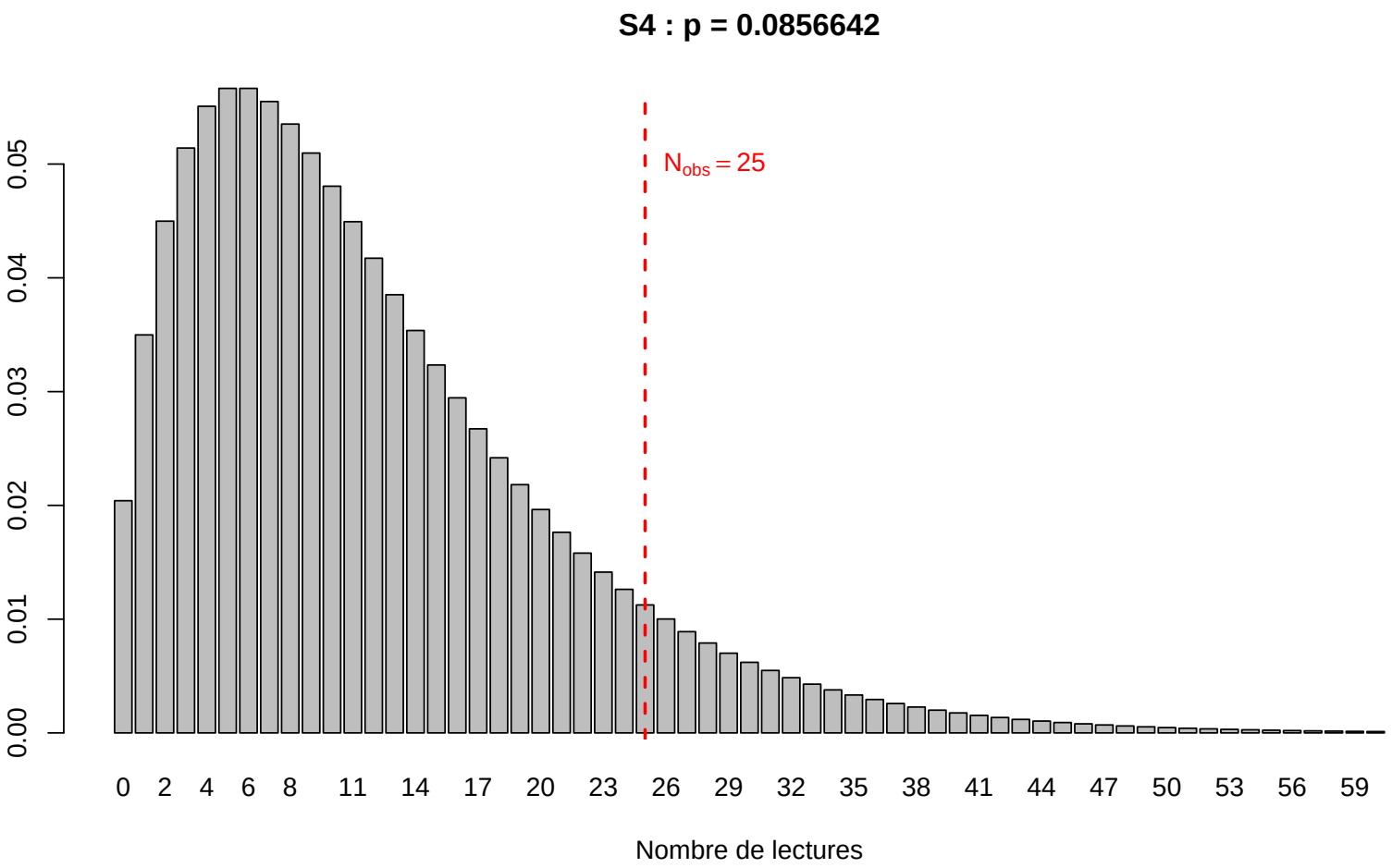
Sur l'exemple précédent



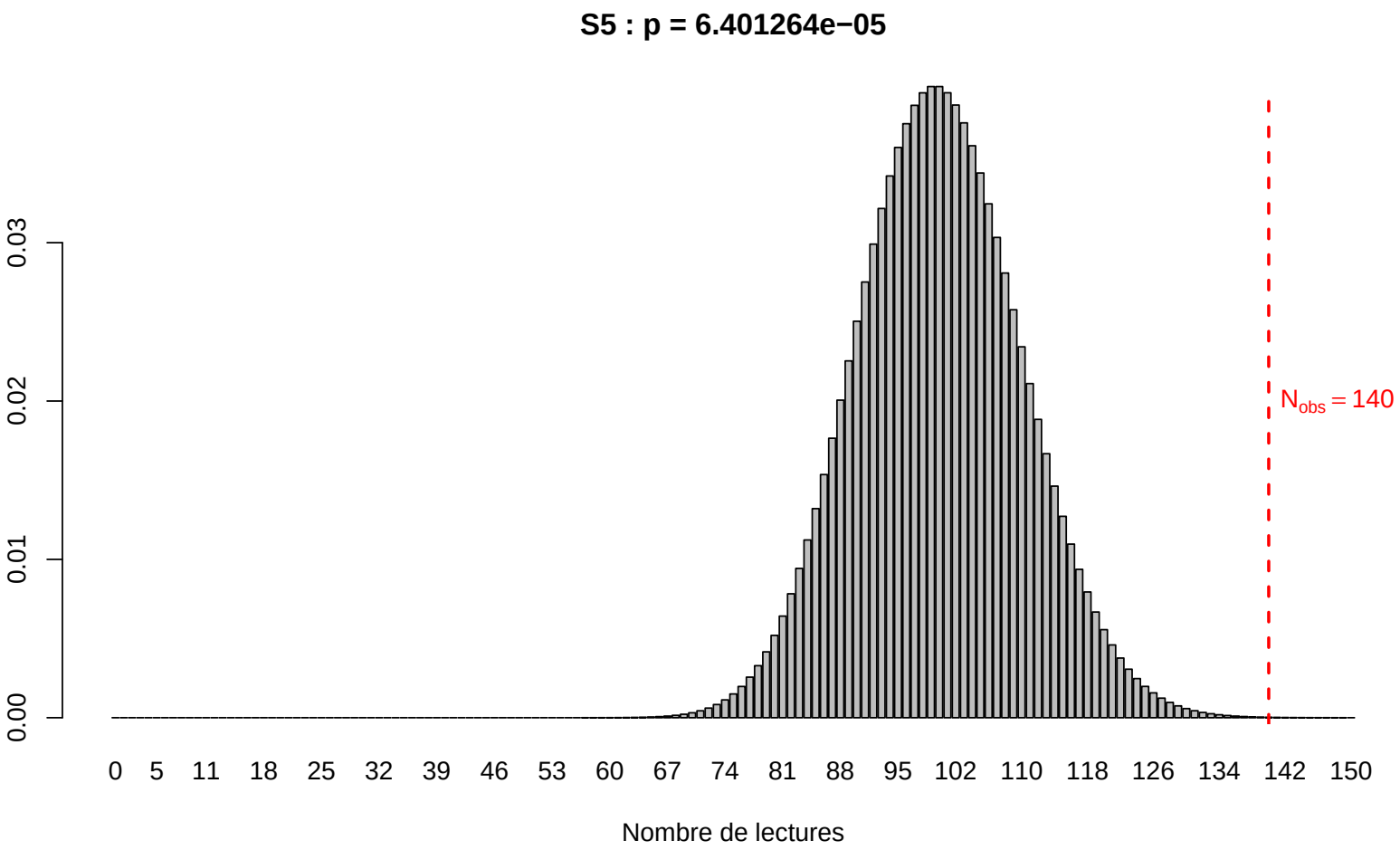
Sur l'exemple précédent



Sur l'exemple précédent



Sur l'exemple précédent



Synthèse 3

Un test statistique (au sens de R. Fisher) est un genre de raisonnement par l'absurde :

- pour démontrer l'existence d'un effet du traitement, on suppose le contraire

$$H_0 : \text{le traitement n'a aucun effet}$$

- on effectue des mesures sur la condition contrôle pour déterminer la distribution d'une variable aléatoire
- on effectue une mesure sur la condition traitement
- on démontre que sous l'hypothèse H_0 ce dernier résultat est très improbable
- ce qui amène une contradiction et prouve l'existence d'un effet

La p-valeur est la probabilité d'obtenir dans le contrôle des résultats plus extrêmes que dans le traitement.

Interprétation pratique de la p-valeur

Comment décider si l'on a la preuve suffisante pour l'existence d'un effet ?

- la p-valeur doit être "suffisamment petite" $\Rightarrow$ choisir un seuil
- difficile de fournir un choix "tout terrain"
- des résultats théoriques montrent qu'un choix de 0.001 n'est pas complètement déraisonnable

Il existe deux façons de se tromper dans cette décision :

- décider que H_0 est fausse alors qu'elle est vraie
  - taux de faux positifs
  - **erreur de type I**
- décider que H_0 est vraie alors qu'elle est fausse
  - taux de faux négatifs
  - **erreur de type II**

Remarque sur les types d'erreur

Les deux types d'erreur n'ont pas les mêmes conséquences

- type I $\rightsquigarrow$ publication erronnée
- type II $\rightsquigarrow$ pas de publication

En pratique,

- on fixe le taux admissible d'erreur de type I (bas de préférence)
- et on essaie de minimiser le taux d'erreur de type II
- assimilable au principe de précaution

Plan

Contexte

Supposons maintenant que l'on réalise le test :

- sur de (très) nombreuses régions n
- dont une écrasante majorité ne présente pas d'effet.
- On pose α le seuil de p-valeur pour le test.
- Le taux d'erreur de type I du test est égal à α
  - puisque α est la probabilité de commettre une erreur de type I lorsqu'il n'y a pas d'effet
- Parmi les régions détectées comme présentant un effet il y aura :
  - les vrais positifs
  - et les faux positifs

Question

Quelle est la probabilité de ne commettre aucune erreur de type I en réalisant les n tests ?

Aucune erreur de type I sur n tests ?

$\mathbb{P}[\text{erreur de type I sur 1 test}] = \alpha$
$\mathbb{P}[\text{pas d'erreur de type I sur 1 test}] = 1 - \alpha$
$\mathbb{P}[\text{pas d'erreur de type I sur 2 tests}] = (1 - \alpha) \times (1 - \alpha)$
$\mathbb{P}[\text{pas d'erreur de type I sur n tests}] = (1 - \alpha)^n$

Or

$$(1 - \alpha)^n \xrightarrow[n \to +\infty]{} 0$$

Soit typiquement :

$$(1 - 0.001)^{100} \approx 0.9047921$$
$$(1 - 0.001)^{1000} \approx 0.3676954$$
$$(1 - 0.001)^{1000} \approx 4.517335.10^{-5}$$

il faut ajuster α en fonction du nombre de tests effectués

Correction de Bonferroni

Objectif

après n tests, la probabilité de commettre au moins une erreur de type I est α.

- rappel

$$\mathbb{P}[\text{au moins une erreur de type I}] = 1 - (1-\alpha)^n$$

- pour x "petit", on a l'approximation

$$(1-x)^n \approx 1 - nx$$

- d'où pour $\alpha \lll \frac{1}{n}$

$$\mathbb{P}[\text{au moins une erreur de type I}] \approx n\alpha$$

Correction de Bonferroni

Principe

pour n tests, on choisit un seuil $\frac{\alpha}{n}$.

- en conséquence la probabilité qu'une au moins des régions soit un faux-positif est α
- stratégie très simple ...
- mais très conservative/stricte : seuls les régions "évidentes" risquent d'être acceptés

Contrôle du taux de faux positif

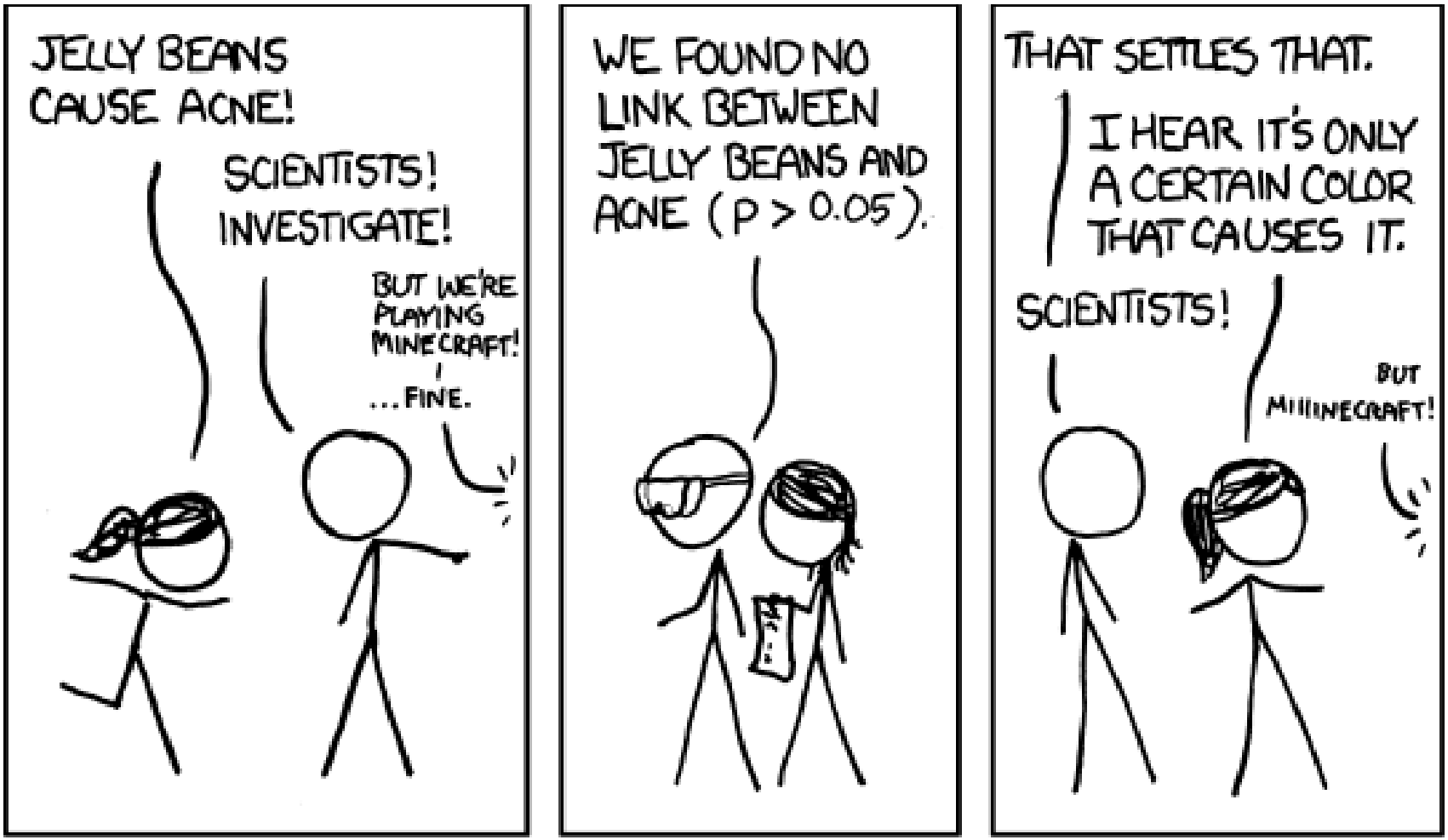
Une stratégie alternative

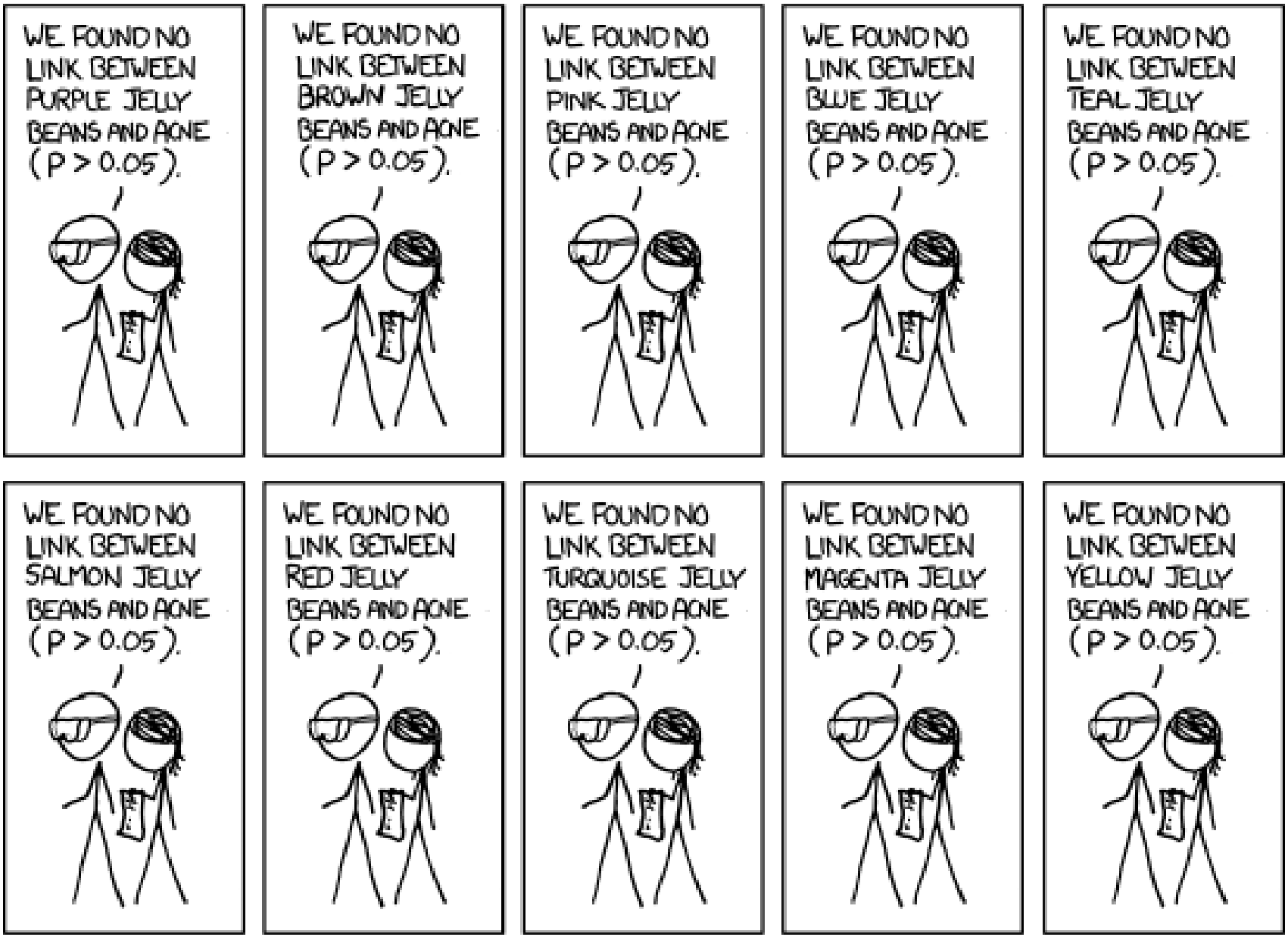
choisir le seuil α de façon qu'après n tests, on trouve dans les régions détectées une fraction x de faux-positifs.

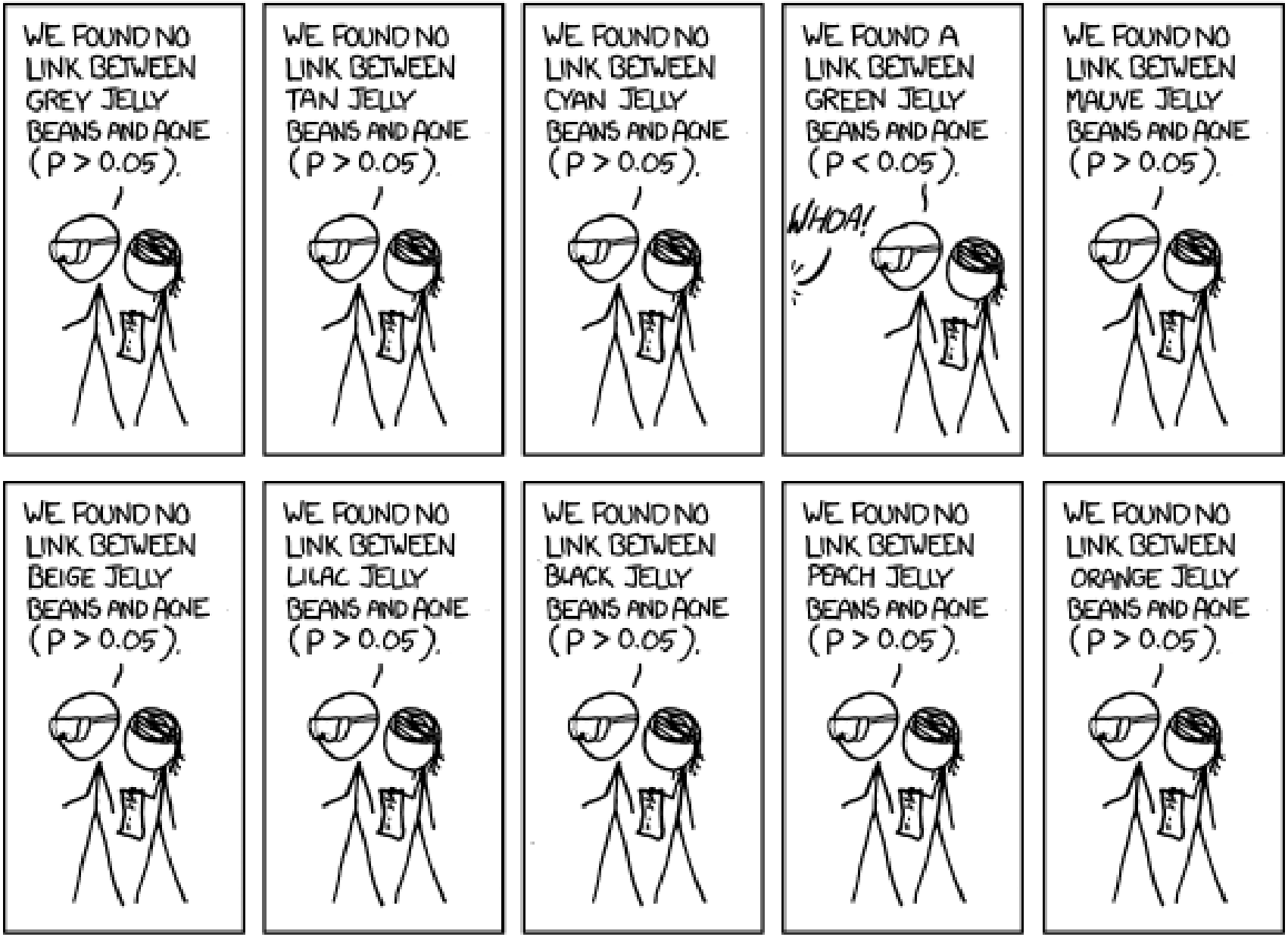
- *a.k.a. False Discovery Rate (FDR)*
- interprétation : en acceptant les régions avec une FDR inférieure à x, on obtient en moyenne une fraction x de faux-positifs parmi les régions acceptées
- contrôle plus raisonnable du nombre de faux positifs
- stratégie très populaire dans les logiciels
- mais plus compliquée à mettre en place (mais pas pour l'utilisateur)

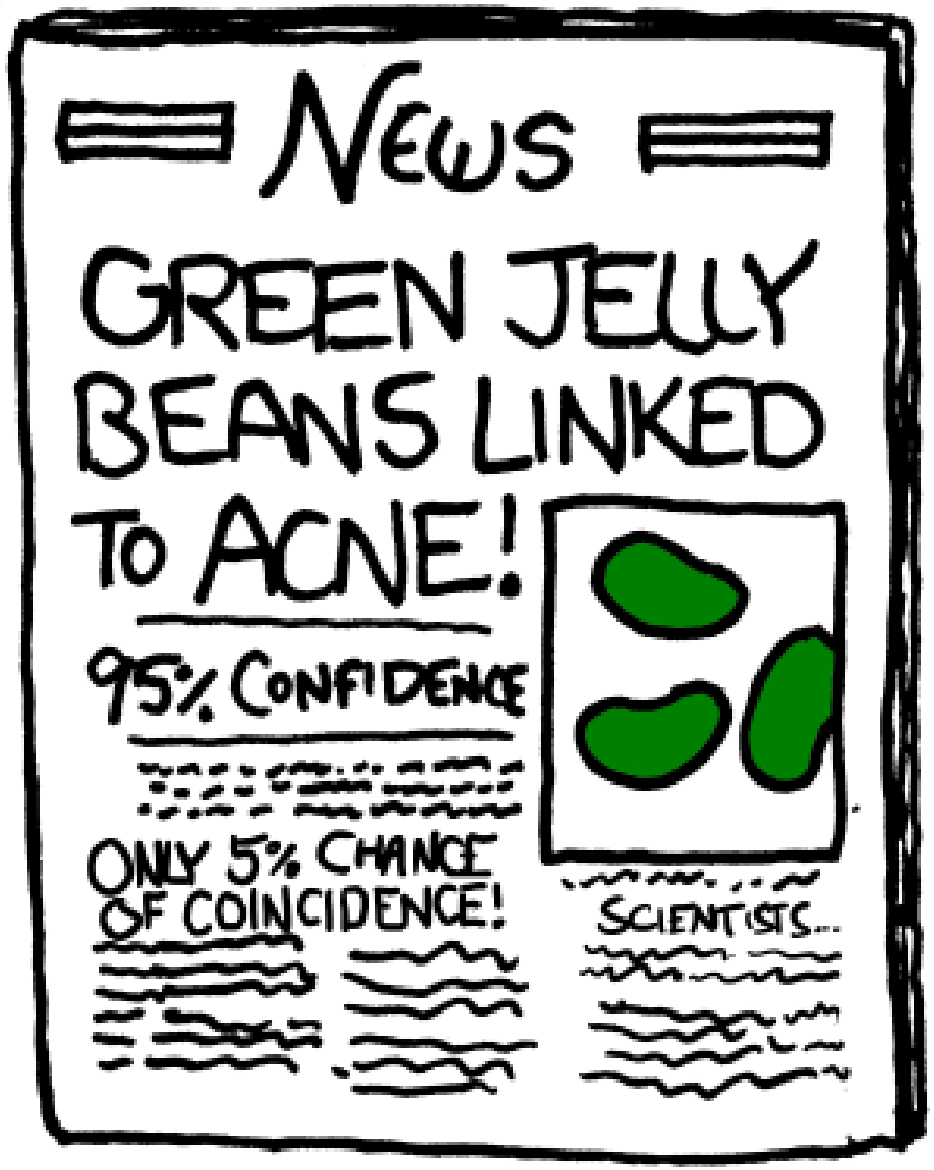
Synthèse 4

- Le filtre par p-valeur est insuffisant/trop optimiste lorsque l'on fait plusieurs tests.
- La correction de Bonferroni contrôle la probabilité de commettre au moins une erreur.
- Le calcul de la FDR permet de contrôler la proportion de faux-positifs dans la liste finale.
- En pratique un filtre FDR $< 1, 5$ ou 10% est un bon début.


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
... FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE ($p > 0.05$).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!


WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).


WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).


NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...