Puissance d'un test

Annette Corpart et Nelly Lassalle

Objectifs des tests de validité d'hypothèse :

- déterminer si la statistique observée dans **un échantillon est conforme à une valeur de référence** supposée connue dans la population, ceci afin de prendre des décisions concernant l'ensemble de la population.
- évaluer le **risque d'erreur pris en prenant la décision** de conformité ou pas de l'échantillon.

Deux erreurs dans la prise de décision :

En prenant une décision, on est confronté à deux types d'erreur :

- « on rejette une hypothèse vraie », c'est l'erreur de première espèce (notée α). Si par exemple, on retient 5% pour seuil de risque, alors il y a à peu près 5 chances sur 100 pour qu'on rejette l'hypothèse quand elle aurait dû être acceptée ; on est confiant à peu près à 95% d'avoir pris la bonne décision.
- « on accepte une hypothèse fausse », c'est l'erreur de seconde espèce (notée β).

Ainsi quatre situations sont possibles dans toute prise de décision :

réalité / décision du test	H_0 est vraie	H_0 est fausse
H_0 est acceptée	Bonne décision Niveau de confiance $1 - \alpha$	Mauvaise décision (acceptation à tort) Erreur β de seconde espèce
H_0 est rejetée	Mauvaise décision (rejet à tort) Erreur α de première espèce	Bonne décision Puissance du test $1 - \beta$

Pour que les tests de conformité soient efficaces, ils doivent minimiser les erreurs de décision. La plupart du temps, les erreurs des deux types n'ont pas la même importance de sorte que l'on essaie de limiter la plus grave. En pratique, on connait α mais on ne sait pas calculer β (il faudrait connaître la loi de probabilité de la variable étudiée sous l'hypothèse alternative H_1).

Attention, lorsque la valeur observée appartient à l'intervalle d'acceptation (c'est-à-dire lorsque les résultats expérimentaux ne sont pas contradictoires avec l'hypothèse nulle), il convient de s'exprimer avec beaucoup de précaution : on ne peut pas affirmer « *l'hypothèse nulle est vraie* » ; il faut dire « *on admet l'hypothèse nulle car rien ne permet de la rejeter* ».
On adopte ainsi une démarche qui s'apparente à la démarche scientifique qui consiste à admettre une théorie jusqu'à la preuve de son échec. Lorsque l'on dit « *admettre* », on ne signifie pas que la théorie est vraie mais qu'elle rend compte pour l'instant - jusqu'à plus ample informé - des expériences.
Exemples : la mécanique générale admise jusqu'à la théorie de la relativité, la mécanique céleste.

Les tests ne sont pas faits pour « démontrer » H_0, mais pour la rejeter .

Que se passe-t-il si on est effectivement dans le cadre de l'hypothèse alternative ? Quel risque prend-on de rejeter H_0 ? La puissance de la décision est sa capacité à détecter les écarts par rapport à l'hypothèse nulle.

On appelle puissance d'un test *P* la probabilité de rejeter l'hypothèse nulle alors qu'elle est fausse.

La valeur complémentaire à 1 de cette puissance, c'est-à-dire la probabilité de ne pas rejeter l'hypothèse nulle alors qu'elle est fausse, est le risque de seconde espèce β : on a donc $P = 1 - \beta$.

La courbe de puissance du test est la représentation graphique de cette probabilité quand le paramètre testé (une moyenne μ ou une proportion p) varie.
Cette fonction « puissance du test » qui associe à la valeur réelle de μ (ou de p), la probabilité de rejeter avec raison H_0 (définie par μ_0 ou p_0) est une fonction décroissante sur $]-\infty ; \mu_0]$ (ou sur $[0 ; p_0]$) et croissante sur $[\mu_0 ; +\infty[$ (ou sur $[p_0 ; 1[$) dont la limite à l'infini (ou en 1) est 1. Le test est de qualité d'autant meilleure que la croissance de la fonction est plus rapide.

Les calculs de puissance en statistique se font généralement avant que l'expérience ne soit menée. En effet, la puissance statistique consentie permet de calculer le nombre de sujets à inclure dans une étude. En général, on fixe la puissance désirée, le risque de première espèce et les paramètres associés aux groupes pour obtenir le nombre de sujets nécessaire à l'étude. On s'en sert donc principalement pour estimer le nombre d'observations nécessaires pour que l'expérience ait la qualité statistique requise.

Premier exemple *(test unilatéral relatif à une proportion)*

Un examinateur doit faire passer une épreuve de type QCM à des étudiants. Ce QCM est constitué de n questions indépendantes. Pour chaque question, il y a trois réponses possibles dont une seule correcte.

Hypothèses : on suppose qu'il y a deux sortes d'étudiants :
Hypothèse nulle H_0 : l'étudiant n'a pas travaillé et répond au hasard : il a alors une chance sur 3 d'avoir une réponse juste. Probabilité de réussite $p = 1/3$
Hypothèse alternative H_1 : l'étudiant a travaillé : il a davantage de chance de donner une bonne réponse à chaque question mais sa probabilité de réussite est inconnue : $p > 1/3$

Pour déclarer l'étudiant « reçu », l'examinateur détermine une valeur k telle que :
- si le nombre de réponses correctes est supérieur ou égal à k , l'étudiant est reçu.
- si le nombre de réponses correctes est strictement inférieur à k , l'étudiant est recalé.

A l'issue des résultats de l'épreuve, quatre cas sont possibles[1] :

(1) l'étudiant n'a pas travaillé et il est recalé

(2) l'étudiant a travaillé et il est reçu

(3) l'étudiant n'a pas travaillé et il est reçu

(4) l'étudiant a travaillé et il est recalé

Si $n = 20$
Considérons la variable aléatoire X égale au nombre de réponses correctes parmi les 20, pour un étudiant choisi au hasard. X suit la loi binomiale $\mathcal{B}(20 ; p)$ avec p : probabilité de réussite de l'étudiant.

On a deux risques d'erreur correspondant aux cas (3) et (4)[2] :
- l'erreur α de première espèce : on rejette l'idée que l'étudiant n'a pas travaillé alors que c'est vrai (c'est le « risque professeur »). On a $\alpha = P(X \geq k)$ où X suit la loi binomiale $\mathcal{B}(20 ; 1/3)$.
- l'erreur β de seconde espèce : on pense que l'étudiant n'a pas travaillé alors que c'est faux (c'est le « risque étudiant »). On a $\beta = P(X < k)$ où X suit la loi binomiale $\mathcal{B}(20 ; p)$. La puissance du test, $P = 1 - \beta$, est la probabilité pour un étudiant qui a travaillé d'être reçu.

L'examinateur cherche à contrôler ces deux erreurs α et β. Il estime la probabilité de réussite d'un étudiant qui a travaillé à 0,6 (hypothèse alternative H_1). Il obtient alors les risques suivants :
Pour $k = 10$: $\alpha = P(X \geq 10) \approx 0{,}09$; $\beta = P(X < 10) \approx 0{,}13$; $P = 0{,}87$
Pour $k = 11$: $\alpha = P(X \geq 11) \approx 0{,}04$; $\beta = P(X < 11) \approx 0{,}24$; $P = 0{,}76$
Pour $k = 12$: $\alpha = P(X \geq 12) \approx 0{,}01$; $\beta = P(X < 12) \approx 0{,}40$; $P = 0{,}60$

[1] Voir fichier excel « simulation du QCM »
[2] Voir fichier géogébra « Les 2 erreurs dans le test du QCM »

Si $n = 100$

L'examinateur jugeant la puissance du test insuffisante, décide, pour diminuer le risque de seconde espèce sans augmenter le risque de première espèce, de poser 100 questions.

On considère maintenant la variable aléatoire F égale à la fréquence de réponses correctes parmi les 100, pour un étudiant choisi au hasard. Le nombre n étant suffisamment grand, on admet que F suit la loi normale de moyenne p (probabilité de réussite de l'étudiant) et d'écart-type $\sqrt{\dfrac{p(1-p)}{100}}$.

Sous H_0, F suit la loi normale de paramètres $\dfrac{1}{3}$ et $\sqrt{\dfrac{1\times 2}{3\times 3\times 100}} \approx 0{,}047$

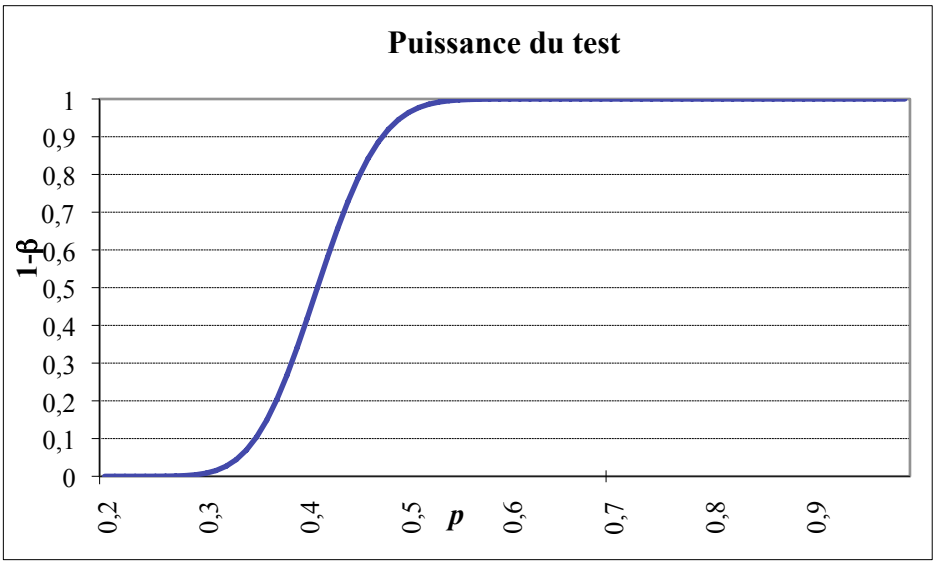
En prenant $\alpha = 0{,}05$, $P(F \geq c) = 0{,}05$ donne comme valeur critique $c = 0{,}41$.

La puissance du test est donnée par le calcul suivant :

$P = 1 - \mathrm{P}(F < c)$ où F suit la loi normale de paramètres p et $\sqrt{\dfrac{p(1-p)}{100}}$.

Construction de la courbe de puissance du test avec un tableur.
On peut présenter les calculs de la façon suivante :
En A1 : p En A2 : 0,20 En A3 : 0,21 ……
En B1 : $1-\beta$ En B2 : =1-LOI.NORMALE(0,41;A2;RACINE(A2*(1-A2)/100);1)
Puis construire la courbe de la puissance du test en fonction de la probabilité p.

	A	B
1	p	1-β
2	0,2	7,60496E-08
3	0,21	4,54718E-07
4	0,22	2,25215E-06
5	0,23	9,4614E-06
6	0,24	3,43867E-05
7	0,25	0,000109925
8	0,26	0,000313445
9	0,27	0,000806789
10	0,28	0,0018938
11	0,29	0,004089886
12	0,3	0,008188654
13	0,31	0,015301623
14	0,32	0,02684315
15	0,33	0,044437925
16	0,34	0,069743892
17	0,35	0,104206402
18	0,36	0,148783124
19	0,37	0,203695673
20	0,38	0,268266499
21	0,39	0,340886034
22	0,4	0,419128243
23	0,41	0,5
24	0,42	0,580280151
25	0,43	0,656885765
26	0,44	0,727200239
27	0,45	0,789310248
28	0,46	0,842121538
29	0,47	0,885350361
30	0,48	0,919411105
31	0,49	0,945236205
32	0,5	0,964069681
33	0,51	0,977271462
34	0,52	0,986158942
35	0,53	0,991898968
36	0,54	0,995451141
37	0,55	0,997554344
38	0,56	0,998743767
39	0,57	0,999384996
40	0,58	0,999713808
41	0,59	0,999873782



Deuxième exemple *(test bilatéral relatif à une proportion)*

Une variété de souris présente des cancers spontanés avec une proportion constante dans la population (connue) $p = 20\,\%$. On se demande si un changement d'alimentation modifie cette proportion (en l'augmentant ou en la diminuant). Pour répondre à cette question on procède à une expérience sur 100 souris et il s'agira, au vu de la fréquence observée f d'animaux cancéreux, de dire si le changement est actif.

Hypothèses en présence :
- H_0 : le changement est inactif ($p = 0{,}2$)
- H_1 : le changement est actif ($p \neq 0{,}2$) .

Intervalle d'acceptation :
On considère la variable aléatoire F_n qui à tout échantillon de n souris associe son pourcentage d'animaux malades.
Si on prend $\alpha = 0{,}05$ et $n = 100$ souris, on obtient l'intervalle d'acceptation I = [0,12 ; 0,28]
Cela signifie que si $p = 0{,}2$ (proportion supposée « vraie »), 95 % des valeurs des fréquences observées sur 100 souris appartiendront à l'intervalle [0,12 ; 0,28].

On adopte alors la **règle de décision** suivante :
- si la fréquence observée de souris cancéreuses parmi les 100 traitées appartient à cet intervalle, on considère que cette valeur est compatible avec les fluctuations d'échantillonnage et l'activité du changement n'est pas prouvée.
- si la fréquence observée n'appartient pas à cet intervalle, le changement sera considéré comme actif et on décidera le rejet de H_0.

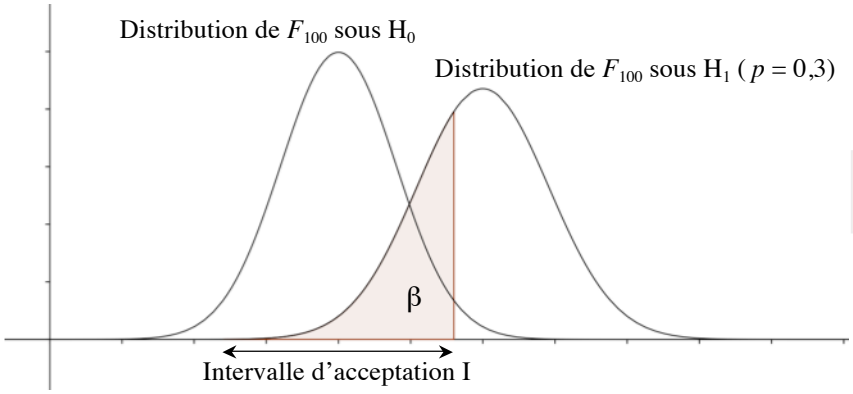
Si l'activité du changement n'est pas démontrée (compte tenu de l'expérience effectuée) peut-on penser qu'une autre expérience, plus complète par exemple, puisse montrer cette efficacité si elle existe ?
Effectivement, l'aptitude d'un test à rejeter l'hypothèse nulle alors qu'elle est fausse est limitée. On va calculer la puissance du test.

Le calcul de la puissance est assez complexe car l'hypothèse alternative est vague. Pour contourner cette difficulté, on considère le cas d'une hypothèse alternative « plus fine ». Par exemple, supposons que l'hypothèse H_1 soit $p = 0{,}3$, l'hypothèse H_0 restant inchangée, c'est-à-dire $p = 0{,}2$.
Dans ces conditions, il est possible de calculer la distribution de la proportion observée, non plus seulement sous l'hypothèse nulle, mais également sous l'hypothèse alternative. On obtient :

• sous l'hypothèse nulle ($p = 0{,}2$), F_n suit la loi normale de moyenne 0,20 et d'écart-type $\sqrt{\dfrac{0{,}20 \times 0{,}80}{n}}$.

• sous l'hypothèse alternative ($p = 0{,}3$), F_n suit la loi normale de moyenne 0,30 et d'écart-type $\sqrt{\dfrac{0{,}30 \times 0{,}70}{n}}$.

La figure ci-dessous présente les deux distributions correspondantes, pour $n = 100$.



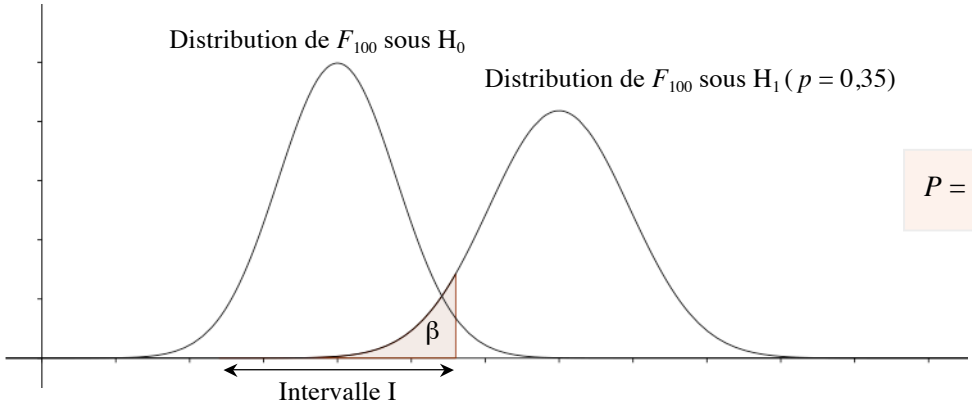
$P = 1 - 0{,}33 = 0{,}67$

La valeur de l'aire colorée est égale à la probabilité que la valeur observée appartienne à l'intervalle d'acceptation du test sachant qu'elle est issue d'une distribution associée à H_1 : c'est le risque de seconde espèce β, le complémentaire à 1 de la puissance du test.

D'où : $P = 1 - P(0{,}12 \leq F_{100} \leq 0{,}28)$ où F_{100} suit la loi normale de moyenne 0,30 et d'écart-type 0,046.

La puissance du test vaut 1 – 0,33 = 0,67 : cela signifie qu'on aura un peu plus de « 6 chances sur dix » seulement de rejeter l'hypothèse $p = 0{,}2$ lorsque p sera égal à 0,3. Autrement dit, environ presque 4 fois sur dix, on sera incapable de détecter que p vaut 0,3 et non 0,2.

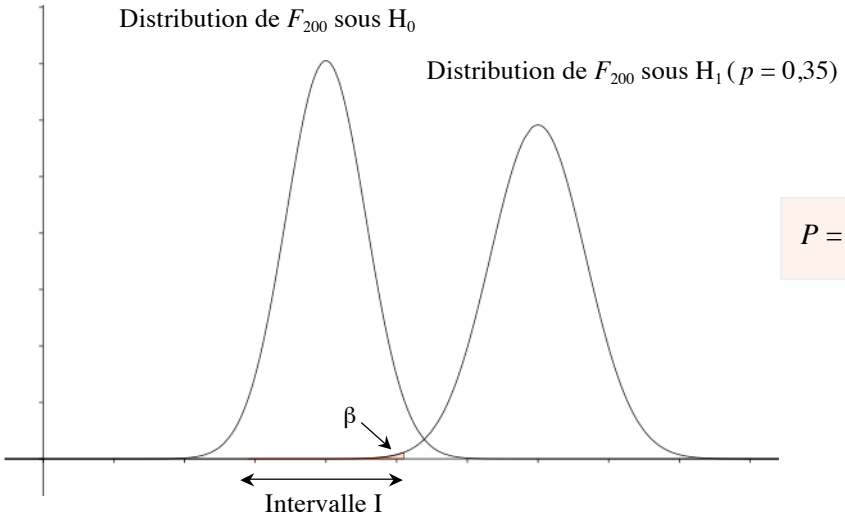
Par ailleurs, on perçoit que plus les hypothèses H_0 et H_1 sont contrastées (par exemple les hypothèses $p = 0{,}2$ et $p = 0{,}35$), plus les distributions de F_n sous ces deux hypothèses sont « éloignées », et plus la puissance est grande.



$P = 1 - 0{,}07 = 0{,}93$

$P = 1 - \mathrm{P}(0{,}12 \leq F_{100} \leq 0{,}28)$ où F_{100} suit la loi normale de moyenne 0,35 et d'écart-type $\sqrt{\frac{0{,}35 \times 0{,}65}{100}} \approx 0{,}048$.

La figure ci-dessous reproduit les mêmes conditions, mais avec une valeur de n plus grande. Autrement dit le même test est mis en œuvre, mais sur un effectif d'échantillon plus important. On constate que le risque de seconde espèce est très faible : **la puissance d'un test augmente avec la taille de l'échantillon.**



$P = 1 - 0{,}004 = 0{,}996$

$P = 1 - \mathrm{P}(0{,}14 \leq F_{200} \leq 0{,}26)$ où F_{200} suit la loi normale de moyenne 0,35 et d'écart-type $\sqrt{\frac{0{,}35 \times 0{,}65}{200}} \approx 0{,}034$.

Remarques :

- On peut tracer point par point la courbe de la fonction « puissance du test »[3] : en abscisse, les valeurs de p ; en ordonnée les valeurs de $P = 1 - \beta$.
- Lorsque l'hypothèse nulle n'est pas rejetée, on peut toujours dire que c'est un **manque de puissance du test** car on estime que H_0 est « sans doute » fausse. On peut penser qu'avec un plus grand nombre d'individus, H_0 serait rejetée. Cela justifie l'expression « l'activité du changement n'est pas démontrée ».
- L'analyse de la puissance statistique d'un test et l'estimation de la taille d'échantillon optimale constituent des aspects essentiels des plans d'expériences, car sans ces calculs, la taille de l'échantillon pourrait être insuffisante ou inutilement grande. Si la taille de l'échantillon est trop faible, l'expérience manquera de précision pour répondre avec fiabilité aux questions posées. Si la taille de l'échantillon est au contraire trop grande, ce seront des ressources et du temps gaspillés pour un gain minime.

Des expériences bien conçues doivent garantir une puissance suffisante pour pouvoir détecter des écarts raisonnables par rapport à l'hypothèse nulle. Sans quoi, l'expérience même risque de présenter peu d'intérêt.

[3] Voir fichier géogébra « Puissance du test sur les cancers de souris »

Troisième exemple *(test unilatéral relatif à une moyenne)*

On s'intéresse à un test pour mesurer la consommation maximale en oxygène d'un individu dans une population âgée. Pour un groupe de contrôle, il a été montré que les mesures suivent une loi normale dont l'espérance mathématique est de l'ordre de $\mu = 25{,}5$ (ml/kg/min) et l'écart-type $\sigma = 6$ (ml/kg/min).
On pense qu'une population de malades (Parkinson) doit avoir des capacités cardio-respiratoires plus limitées. On souhaite donc tester si dans un tel groupe la moyenne μ est plus faible.
Le principe du test est donc de décider entre deux hypothèses :

- l'hypothèse nulle notée H_0 : $\mu = 25{,}5$ (absence d'effet de la maladie)
- l'hypothèse alternative notée H_1 : $\mu < 25{,}5$ (existence de l'effet).

On considère la variable aléatoire $\overline{X}_n$ qui à tout échantillon de n personnes associe sa consommation maximale moyenne en oxygène.

Pour un seuil de risque $\alpha = 0{,}05$ et pour une expérience portant sur $n = 15$ personnes :

$\overline{X}_{15}$ suit la loi normale de moyenne 25,5 et d'écart-type $\frac{6}{\sqrt{15}}$;

$P(\overline{X}_{15} < c) = 0{,}05$ donne comme valeur critique $c = 22{,}95$.

La question est de savoir quel risque on prend si on rejette H_0. Il faut choisir dans quelle mesure on s'écarte de l'hypothèse nulle : c'est une décision qui se prend à partir de considérations scientifiques. Le statisticien ne peut ici se substituer au praticien. Il lui demande en particulier à partir de quelle valeur un effet constitue une différence scientifiquement significative.
On supposera qu'il répond qu'à partir de 23,5 l'effet peut être considéré comme important.

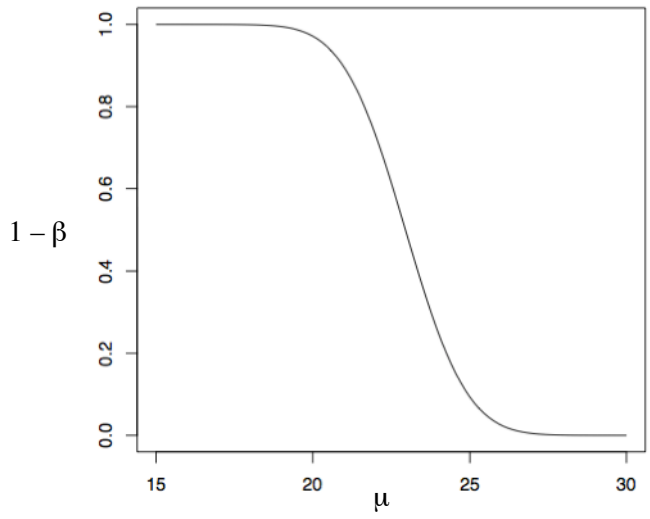
On va donc rejeter l'hypothèse nulle si $\overline{X}_{15} < 22{,}95$. Quelle est la probabilité de cet événement lorsque nous sommes dans le cadre de l'hypothèse alternative avec $\mu = 23{,}5$?

$\overline{X}_{15}$ la loi normale de moyenne 23,5 et d'écart-type $\frac{6}{\sqrt{15}}$ et $P(\overline{X}_{15} < 22{,}95) = 0{,}36$

On constate que l'on a une très faible chance de démontrer ce qui nous intéresse : la puissance de ce test n'est pas satisfaisante.

On considère généralement que la puissance doit au moins être égale à 0,80 pour être satisfaisante.

Si on s'intéresse à une alternative où la moyenne est μ, la puissance est donnée par le calcul : $P(\overline{X}_{15} < 22{,}95)$ avec $\overline{X}_{15}$ qui suit la loi normale de moyenne μ et d'écart-type $\frac{6}{\sqrt{15}}$.
On peut donc calculer la puissance du test et obtenir la courbe de puissance du test. La courbe de puissance est représentée dans la figure ci-dessous.
On voit sur le graphique que, pour une puissance "satisfaisante" de 0,80, on obtient $\mu = 21{,}5$.



Quatrième exemple *(test bilatéral relatif à une moyenne)*

La cote X (en mm) d'une pièce produite en très grande quantité suit une loi normale de moyenne 5 et d'écart-type 0,4. En vue de tester l'hypothèse $\mu = 5$, on prélève dans la production (au hasard et avec remise) un échantillon de 100 pièces dont on mesure la cote moyenne. On obtient $\bar{x} = 5{,}05$. La machine est–elle déréglée ?

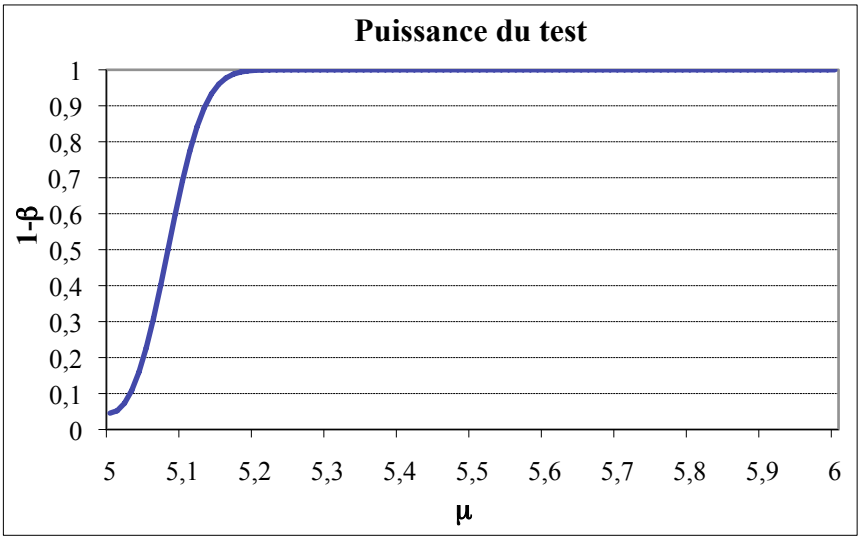
Hypothèses : Hypothèse nulle $H_0 : \mu = 5$
Hypothèse alternative $H_1 : \mu \neq 5$

Intervalle d'acceptation : On note $\overline{X}_{100}$ la variable aléatoire qui à tout échantillon de 100 pièces associe sa cote moyenne. $\overline{X}_{100}$ suit la loi normale de paramètres $\mu = 5$ et $\sigma = 0{,}04$.
En choisissant $\alpha = 5\%$, on obtient l'intervalle $I = [4{,}92 ; 5{,}08]$

La **puissance du test** est donnée par le calcul suivant :
$P = 1 - \beta = 1 - P(4{,}92 \leq \overline{X}_{100} \leq 5{,}08)$ où $\overline{X}_{100}$ suit la loi normale de moyenne μ et d'écart-type 0,04.

Nous allons construire **la courbe de puissance** de ce test avec un tableur.
On peut présenter les calculs de la façon suivante :
En A1 : μ En A2 : 5 En A3 : 5,01 … En A102 : 6
En B1 : $1 - \beta$ En B2 : =LOI.NORMALE(4,92;A2;0,04;1)+1-LOI.NORMALE(5,08;A2;0,04;1)
Puis construire la courbe de la puissance du test en fonction de la moyenne μ.

	A	B
1	μ	1-β
2	5	0,04550026
3	5,01	0,05228363
4	5,02	0,07301687
5	5,03	0,10862954
6	5,04	0,16000515
7	5,05	0,22720438
8	5,06	0,30877017
9	5,07	0,40138209
10	5,08	0,50003167
11	5,09	0,59871701
12	5,1	0,69146586
13	5,11	0,77337366
14	5,12	0,84134503
15	5,13	0,8943503
16	5,14	0,93319282
17	5,15	0,95994085
18	5,16	0,97724987
19	5,17	0,98777553
20	5,18	0,99379033
21	5,19	0,99702024
22	5,2	0,9986501
23	5,21	0,99942297
24	5,22	0,99976737
25	5,23	0,99991158
26	5,24	0,99996833
27	5,25	0,99998931
28	5,26	0,9999966
29	5,27	0,99999898
30	5,28	0,99999971
31	5,29	0,99999992
32	5,3	0,99999998
33	5,31	1
34	5,32	1
35	5,33	1



Avec ce graphique on lit que, pour avoir une puissance satisfaisante de 0,80, μ doit être supérieure à 5,12.