

Statistiques, séance n°4

Les tests statistiques

EPITA

Avril 2020

Problématique

- On vous formule une « **hypothèse** » sur un (nouveau) procédé, sur la description d'un phénomène, la relation de causes à effet entre des événements,
- Vous ne disposez pas de « démonstration » ou d'« explication scientifique » sur ce qui est avancé
- En revanche, vous pouvez vous livrer à des expériences / des tests sur des échantillons
- Sur la base des observations, vous aboutirez à deux types de conclusions :
 - Le rejet de l'hypothèse
 - L'acceptation de l'hypothèse ~ mais qui ne constitue pas une preuve en soi !!
- Pour cela, vous allez mettre en avant une « **variable de décision** »

Exemple, tiré de l'ouvrage de G. Saporta : les faiseurs de pluie

- Les faiseurs de pluie : avec des nuages d'iodures
- Sans faiseur de pluie : $m=600$, $\sigma=100$ (H_0)
- Avec : $m=650$ (H_1)
- Attention au choix de H_0
- $\alpha = 5\%$
- On accepte H_1 si le résultat de l'expérience est tel que H_0 doit être rejeté

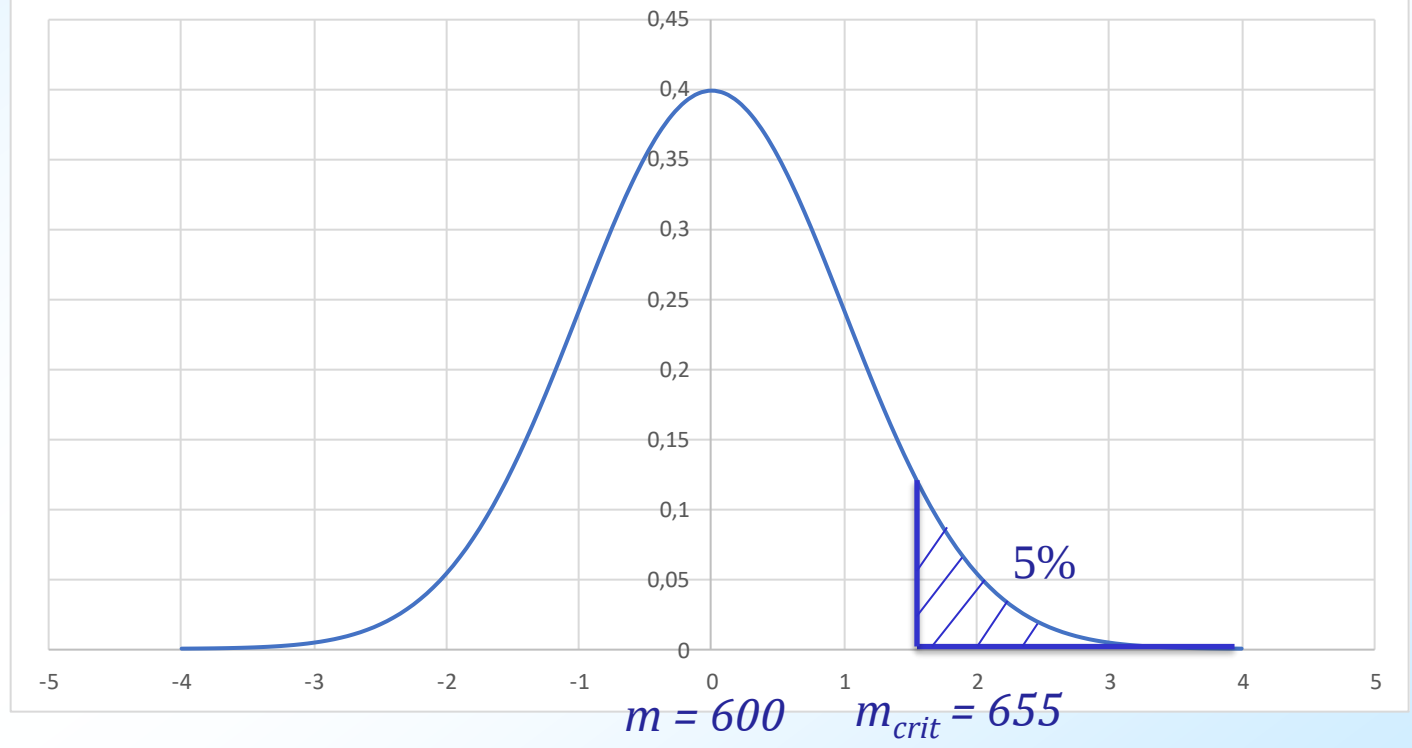
- Ici : $H_0 \cong$ les nuages d'iode n'ont aucun effet
- C'est l' « hypothèse nulle »
- L'hypothèse alternative H_1 : les nuages ont un effet sur l'agriculture
- Il va falloir convaincre les agriculteurs que H_1 est vraie
- Pour cela les résultats de l'expérience doivent se trouver dans une zone où H_0 est improbable
- Avec $\alpha = 5\%$ on va accepter H_1 si la moyenne des hauteurs observée sur les 9 ans est dans une zone improbable pour H_0
- D'où la valeur critique : $m_{crit} = m + \Phi^{-1}(95\%) \frac{\sigma}{\sqrt{9}} = 655$

- $m = 600, \sigma = 100$

- 9 ans => « beaucoup d'observations » => moyenne / Loi Normale !!

- $m_9 = 600, \sigma = \frac{100}{\sqrt{9}} = 33,3$

Densité Loi Normale Standard



$$m + \sigma = 633$$

- Avec nos données (cf. fichier Excel) $m_{constaté} = 610,2$
- On ne rejette pas H_0
- A priori, la recette des faiseurs de pluie ... ne fait pas recette !

- **Test paramétrique : test de certaines hypothèses portant sur un ou plusieurs paramètres du V.A.**
- **Tests non paramétriques**
- **Hypothèse simple : $\theta = \theta_0$**
- **Hypothèse composite : $\theta \in A$, souvent : $\theta < \theta_0$, $\theta > \theta_0$, $\theta \neq \theta_0$**

Tests et erreurs

Vérité Décision	H_0	H_1
H_0	$1 - \alpha$	β
H_1	α	$1 - \beta$

- α : risque de première espèce
- β : risque de deuxième espèce
- $1 - \beta$: puissance du test
- Région critique W , ensemble des valeurs de la (des) variable(s) de décisions tq.

$$\begin{array}{ll} \mathbb{P}(W|H_0) = \alpha & \\ \mathbb{P}(\overline{W}|H_0) = 1 - \alpha & \mathbb{P}(W|H_1) = 1 - \beta \end{array}$$

Démarche (*G. Saporta*) retour 17h10

- Choix de H_0 et H_1
- Détermination de la variable de décision
- Allure de la région critique en fonction de H_1
- Calcul de la région critique en fonction de α
- Calcul éventuel de la puissance $1 - \beta$

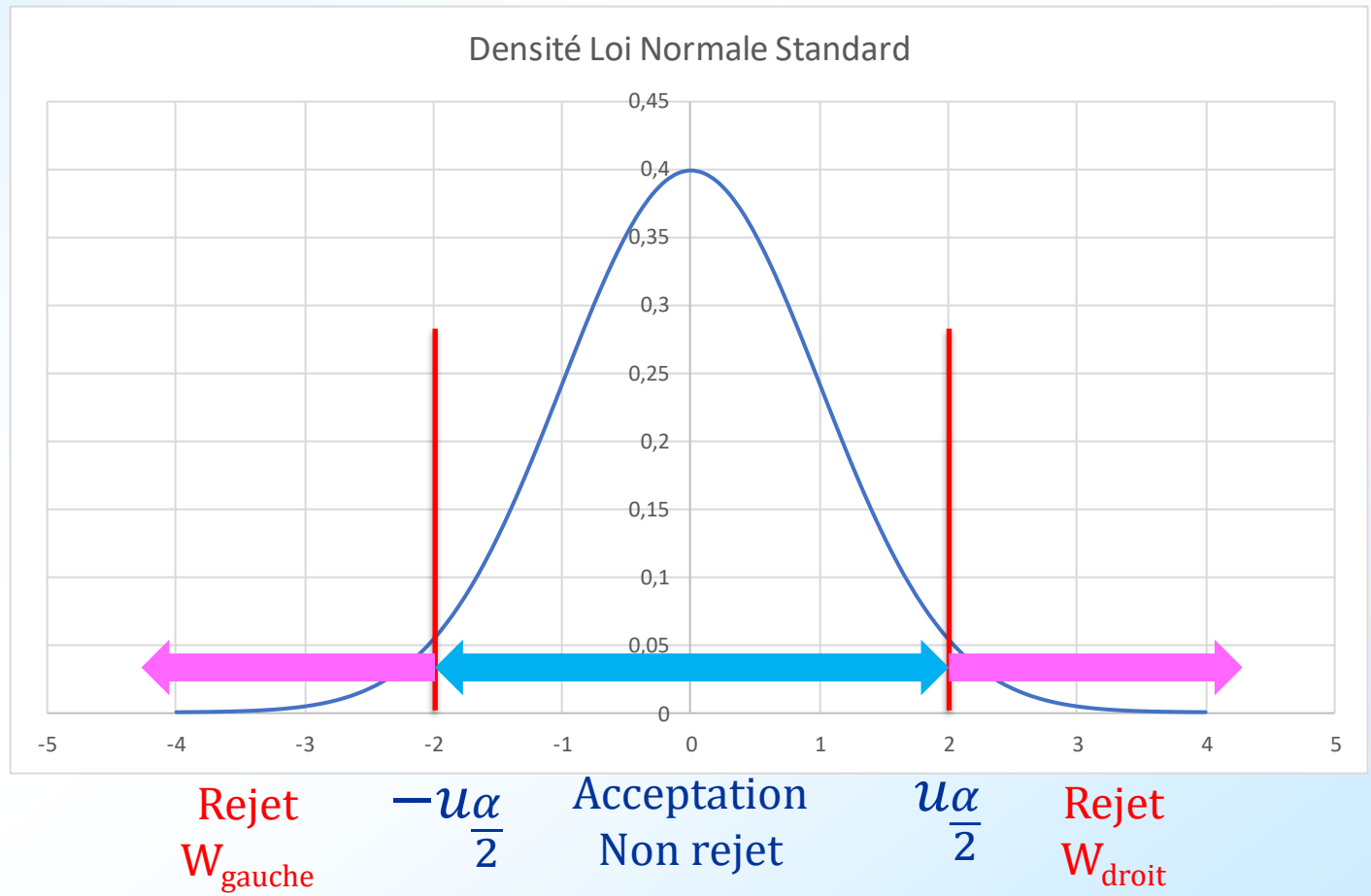
- Calcul de la valeur expérimentale de la variable de décision
- Conclusion rejet ou acceptation de H_0

- Exemple : domaine médical

Premier exemple : test paramétrique

test d'une moyenne μ , σ connu, α fixé

- Dans ce cas, $H_0 : \mu = \mu_0$, $H_1 : \mu \neq \mu_0$
- On « pense que $\mu = \mu_0$ » et il va falloir nous convaincre du contraire !!
- Quand n est grand, $\frac{\sqrt{n}}{\sigma} (\bar{X}_n - \mu_0)$ sous H_0 , suit une loi $\mathcal{N}(0, 1)$
- Variable de décision : $Z_n = \frac{\sqrt{n}}{\sigma} (\bar{X}_n - \mu_0)$
- Comme dans le cours précédent (même raisonnement), on pose : $u_{\frac{\alpha}{2}} = \Phi^{-1}(1 - \frac{\alpha}{2})$
- On rejette H_0 si Z_n est en dehors de l'intervalle $[-u_{\frac{\alpha}{2}}, u_{\frac{\alpha}{2}}]$
- En revanche, si $-u_{\frac{\alpha}{2}} \leq Z_n \leq u_{\frac{\alpha}{2}}$ on ne rejette pas H_0 .



Test du Chi2 ~ *un exemple*

- Un échantillon de n individus est rangé dans k classes, $Cl_1, \dots, Cl_k$.
- On émet l'hypothèse que pour tout i : $\mathbb{P}(Cl_i) = p_i$
- On va tester cette hypothèse,
- Ici, $H_0 : \forall i, \mathbb{P}(Cl_i) = p_i$

- Notre échantillon contient N_i individus dans la classe Cl_i

Les degrés de libertés

- De nombreuses applications existent à cette formulation
- Ici, il y a k classes, mais, si on connaît les $N_1, \dots, N_{k-1}$, alors on connaît aussi N_k
- L'ensemble des valeurs $N_1, \dots, N_k$ a $k - 1$ degrés de libertés
- Si on avait disposé les cases autrement, par exemple
 - dans un tableau rectangulaire (ex des vacances) en faisant une hypothèse d'indépendance de v.a.,
 - En construisant les cases en « ajustant une variable aléatoire » donc en « choisissant les meilleurs » les p_i ,ce nombre aurait été différent.

La variable de décision

- On va demander à ce qu'il y ait au moins 5 éléments dans chaque case, sinon on en regroupe
- La variable de décision est la suivante :
- $Z_n = \chi^2 = \sum_{i=1}^k \frac{(N_i - np_i)^2}{np_i}$
- On montre que, quand n est grand, Z_n suit une loi de χ^2 à $k - 1$ degrés de libertés

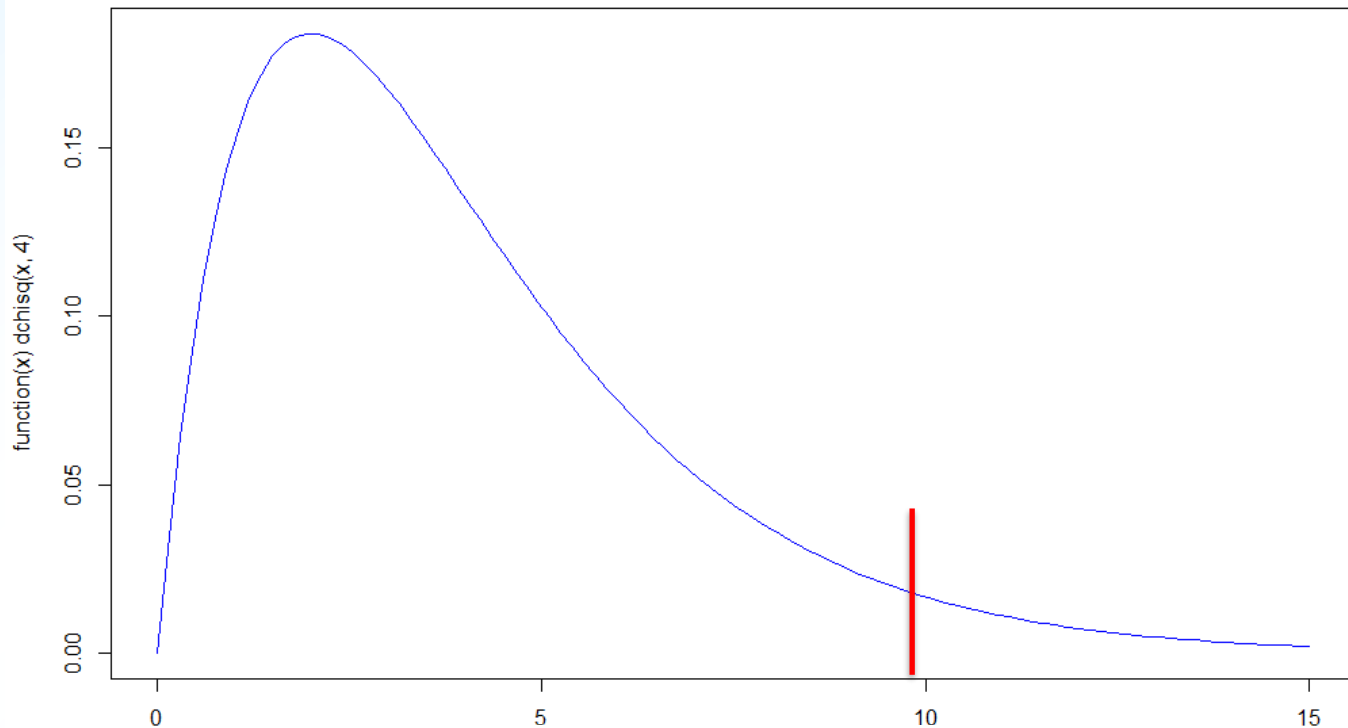
La loi de chi 2

- **En direct sous R**
 - Une simulation
 - La fonction `qchisq` pour notre cours
- **On montre que les distributions de chi2 suivent des loi Gamma,**

Rejet, non rejet

- On rejette l'hypothèse H_0 si Z_n dépasse une valeur, avec R, la valeur en question est

$$qchisq(\alpha, k - 1)$$



Exemple, avec $k=5$ et $\alpha=95\%$ rejet si $Z_n > 9,49$

Nombreuses applications

- **Tests d'indépendances de variables aléatoires**

- Exemple : vacances du deuxième cours
- Précaution avec les degrés de liberté

- **Tests de comportement d'une variable aléatoire**

- On veut tester qu'une variable d'un échantillon respecte une loi donnée,
- On va construire les classes ET ajuster les « meilleurs » p_i en fonction de notre hypothèse de comportement
- Précaution avec les degrés de liberté

-

- **Applications concrètes en médecine, en cryptographie, ...**

Un exemple

- Nous reprenons l'exemple du jeu de pile où face
- Nous supposons que le jeu n'est pas truqué
- Nous avons effectué 1000 lancers
- Nous avons trouvé 460 piles, 540 faces ...
- ... au lieu de 500 / 500 dans une situation idéale
- Nous dans le cas où :

$$n = 1000, p_1 = p_2 = \frac{1}{2}, N_1 = 460, N_2 = 540$$

- On a : $Z_2 = \frac{(N_1 - np_1)^2}{np_1} + \frac{(N_2 - np_2)^2}{np_2} = \frac{(460 - 500)^2}{500} + \frac{(540 - 500)^2}{500} = 6,4$
- Cette variable suit une loi de χ^2 à 1 degré de liberté,
- Si la pièce n'est pas truquée, au seuil 99%, $< 6,63$
- On ne peut pas rejeter H_0
- Remarque : avec $N_1 = 459, N_2 = 541$ on rejette H_0