

Statistiques : Enoncé du TP4
Test d'hypothèses sur un échantillon

On soumet un élève à une série de 20 questions à choix multiples (QCM). On considère ces 20 questions et les réponses données par l'élève comme un échantillon de toutes les questions qu'on pourrait lui poser et des réponses correspondantes de l'élève. Chaque question vient avec la proposition de r réponses dont une seule est correcte ($r = 2, 3, 5, 10$ dans ce qui suit). On modélise la réponse de l'élève à une question donnée par une variable aléatoire de Bernoulli d'espérance p où p caractérise l'élève comme suit :

- $p = \frac{1}{r}$ si l'élève répond au hasard ;
- $p > \frac{1}{r}$ si l'élève répond suivant ses connaissances.

Dans notre modèle on limitera p à deux valeurs possibles : $\frac{1}{r}$ et $\frac{2}{3}$.

On considère dans notre modèle que les questions sont indépendantes entre elles et que l'élève y répond avec un zèle constant. La réponse de l'élève au 20 questions est donc modélisée par une suite de 20 va de Bernoulli indépendantes et identiquement distribuées d'espérance p . Aux 20 réponses (ω) apportées par l'élève on associe la note $S(\omega) = \sum_{1 \leq i \leq 20} X_i(\omega)$ et le taux de réussite $\frac{S(\omega)}{20}$.

1. Simulation d'une copie de l'élève.

a. Pour p un réel entre 0 et 1 donné, l'instruction `bool2s(rand()<p)` simule une va de Bernoulli d'espérance p . Vérifier en prenant $p = \frac{1}{3}$ et en calculant la moyenne des 1000 premières valeurs rendues par cette instruction. Quel est l'intervalle de confiance associée à la moyenne trouvée pour une confiance de 95% ?

Calculer 100 fois de suite cette moyenne et compter le nombre de fois que $\frac{1}{3}$ est dans l'intervalle de confiance. Est-ce compatible avec les 95% ?

b. La somme $S = \sum_{1 \leq i \leq 20} X_i$ est une va suivant la loi binomiale $\mathcal{B}(20, p)$. On a donc pour tout $k \in \{0, \dots, 20\}$

$$P(S = k) = \binom{n}{k} p^k (1 - p)^{n-k} .$$

L'instruction `b=binomial(p,20)` définit un vecteur $b = (b(1), \dots, b(21))$ formée des valeurs $P(S = k)$, k décrivant $\{0, \dots, 20\}$. L'instruction `plot2d3(b)` produit un graphique représentant les valeurs prises par le vecteur b .

On peut simuler les valeurs prises par S par l'instruction `sum(bool2s(rand(1:20)<p))`. Définissez, pour $p = \frac{1}{3}$ un vecteur $s = (s(1), \dots, s(1000))$ formé des 1000 premières valeurs rendues par l'instruction (Définissez s par l'instruction `s=[,]` suivi d'une boucle `for i=1:1000,...`). Tracez en l'histogramme par l'instruction `xset("window",1);histplot([0.5:20],s)` et comparez avec le premier graphique. Qu'obtient-on si on remplace les 1000 itérations par 50 ?

c. L'instruction `sum(b(1:10))` rend la somme des 10 premières valeurs de b donc la probabilité que S soit inférieur ou égal à 10. Définissez un vecteur $F = (F(1), \dots, F(21))$ dont la $k + 1$ -ème valeur est la probabilité que S soit inférieur ou égal à k .

Executer l'instruction `xset("window",2);plot2d2([0:20],F)`. Reconnaissez vous la fonction de répartition de S ?

d. Avec le graphique ou avec le vecteur F , calculez la probabilité que S soit supérieur ou égal à 10. Faites le pour $p = \frac{1}{2}, \frac{1}{3}, \frac{1}{5}$. Comparer avec la proportion de valeurs du vecteur s supérieures ou égales à 10.

e. Pour $p = \frac{1}{2}, \frac{1}{3}, \frac{1}{5}, \frac{1}{10}$ déterminer l'entier k_p minimal tel qu'on ait

$$P(S > k_p) \leq 5\% .$$

On pourra s'aider du tableau suivant

$P(S \leq k)$	$p = \frac{1}{2}$	$p = \frac{1}{3}$	$p = \frac{1}{5}$	$p = \frac{1}{10}$
$k = 0$	0.0000010	0.0003007	0.0115292	0.1215767
$k = 1$	0.0000200	0.0033080	0.0691753	0.391747
$k = 2$	0.0002012	0.0175926	0.2060847	0.6769268
$k = 3$	0.0012884	0.0604465	0.4114489	0.8670467
$k = 4$	0.0059090	0.1515109	0.6296483	0.9568255
$k = 5$	0.0206947	0.2972139	0.8042078	0.9887469
$k = 6$	0.0576591	0.4793427	0.9133075	0.9976139
$k = 7$	0.1315880	0.6614715	0.9678573	0.9995844
$k = 8$	0.2517223	0.8094511	0.9900182	0.9999401
$k = 9$	0.4119015	0.9081042	0.9974052	0.9999928
$k = 10$	0.5880985	0.9623634	0.9994366	0.9999993
$k = 11$	0.7482777	0.9870267	0.9998983	0.9999999
$k = 12$	0.8684120	0.9962754	0.9999848	1.
$k = 13$	0.9423409	0.9991212	0.9999982	1.
$k = 14$	0.9793053	0.9998326	0.9999998	1.
$k = 15$	0.9940910	0.9999749	1.0000000	1.
$k = 16$	0.9987116	0.9999972	1.	1.
$k = 17$	0.9997988	0.9999998	1.	1.
$k = 18$	0.9999800	1.0000000	1.	1.
$k = 19$	0.9999990	1.	1.	1.
$k = 20$	1.	1.	1.	1.

2. Test d'une hypothèse sur p pour une copie donnée.

L'élève ayant rendu sa copie, on veut tester l'hypothèse $p = \frac{1}{r}$ (c'est à dire l'élève a répondu au hasard) dans le cadre de notre modèle. Pour cela on choisit un réel δ et on regarde si le taux de bonnes réponses $\frac{1}{20} \sum_{1 \leq i \leq 20} X_i(\omega)$ est dans l'intervalle $[0, \frac{1}{r} + \delta]$.

a. Dira t-on qu'on rejette l'hypothèse $p = \frac{1}{r}$ si le taux de bonnes réponses est dans l'intervalle $[0, \frac{1}{r} + \delta]$ ou si le taux de bonnes réponses n'est pas dans cet intervalle ?

b. Pour $r = 2$ puis $3, 5, 10$ comment doit on choisir $\delta = \delta_r$ pour que la probabilité de rejeter à tort l'hypothèse $p = \frac{1}{r}$ soit $\leq 5\%$?

c. Pour $r = 2$ puis $3, 5, 10$ et δ_r trouvé ci-dessus, quelle est la probabilité de ne pas rejeter l'hypothèse $p = \frac{1}{r}$ sachant $p = \frac{2}{3}$?

d. Pour $r = 2$ puis $3, 5, 10$ comment faudrait-il choisir δ dans le test défini au début du paragraphe 2 pour que la probabilité de ne pas rejeter l'hypothèse $p = \frac{1}{r}$ sachant que $p = \frac{2}{3}$ soit inférieure à 5% ? Comparer avec les valeurs trouvées dans la question b.

3. Validation du test sur un ensemble de copies.

Les instructions suivantes construisent pour $r = 3$ une matrice copie(1000,20) représentant 1000 copies. Les 500 premières sont construites avec $p = \frac{1}{r}$. Les 500 restantes sont construites avec $p = \frac{2}{3}$. La dernière instruction construit un vecteur s formé du nombre de bonnes réponses de chaque copie.

```
r=3;N=500
```

```
p=[ones(N,1)*1/r;ones(N,1)*2/3] //colonne de longueur 2N
```

```
mp=p;for i=1:19;mp=[mp,p];end //matrice formee de 20 repetitions de p
```

```
copies=bool2s(rand(ones(2*N,20))<mp) //matrice 2Nx20 formee de bernouilli p(i)
```

```
s=copies*ones(20,1) //colonne formee des sommes des lignes de copies
```

a. Pour $r = 3$ et $\delta = \delta_r$, combien de copies parmi les 500 premières sont rejetées à tort par le test

$“\frac{1}{20} \sum_{1 \leq i \leq 20} X_i(\omega)$ est dans l'intervalle $[0, \frac{1}{r} + \delta]”$?

Combien de copies parmi les 500 dernières n'ont pas été rejetées par le test ? Le test est-il satisfaisant ?

Que se passe t-il si on diminue δ ?

*b. Quel est l'intervalle de confiance sur le nombre de copies trouvé ?

c. Recommencez avec $r = 5, 10$. Conclusion ?