UE Probabilités et statistique 1

Ensimag 1A – Examen blanc

Corrigé

Décembre 2025 – Durée 1h

Consignes importantes :

- Ce document **Rmd** constitue votre document de travail. Le devoir doit être **rendu au format PDF**, après avoir complété les espaces réservés aux sections **Réponse.**
- Il est **fortement conseillé** de compiler (`Knit`) le document régulièrement afin de vérifier l'absence d'erreurs.
- Pour valider une **Question** comportant un code R à compléter, il est nécessaire de **modifier l'instruction** `eval = FALSE` en `eval = TRUE` ou de la supprimer.
- Les réponses attendues devront être clairement rédigées, si possible **en gras**, et ne pas être limitées aux codes informatiques à compléter.

Énoncé

On dispose d'un jeu de données `x` contenant les résultats de $\mathbf{n = 94}$ épreuves d'un jeu de pile ("P") ou face ("F"). Le jeu de données est chargé en mémoire de la manière suivante :

```
# charge le jeu de données
x <- scan(file = "pile_ou_face.txt", what = character())
head(x)
```

```
## [1] "F" "P" "P" "P" "F" "P"
```

Question 1

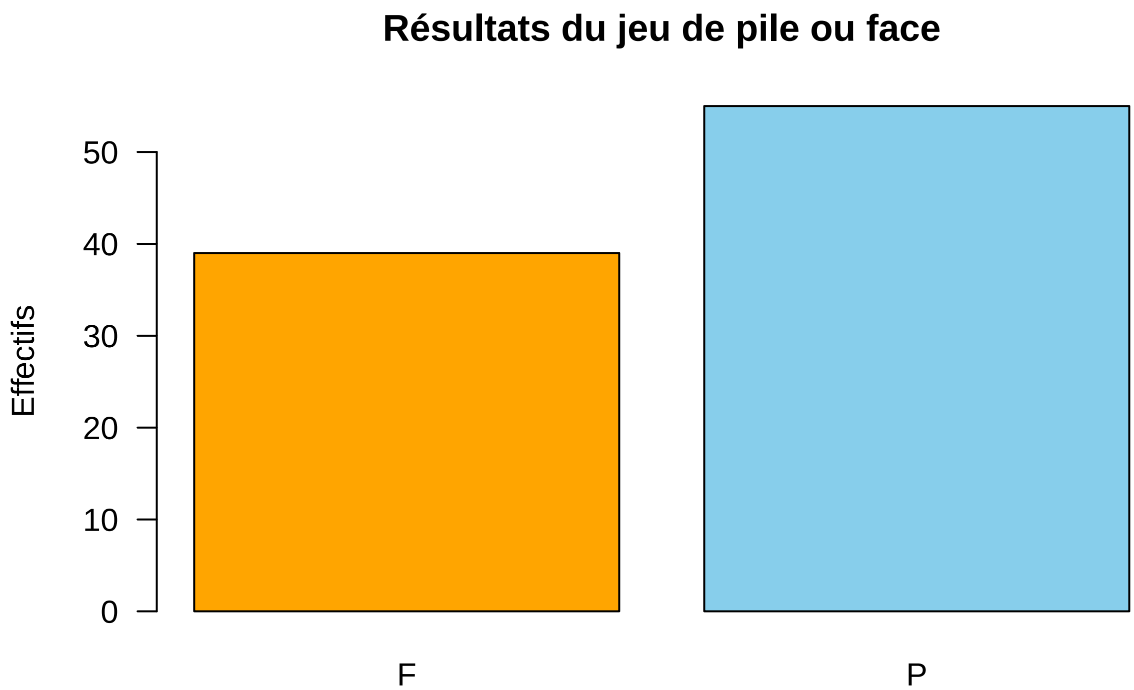
À l'aide de commandes R, calculer le nombre de faces ("F") et de piles ("P") dans le jeu de données. Représenter les effectifs observés à l'aide d'un diagramme en barres.

Réponse. Compléter les codes suivants et décrire brièvement les résultats obtenus.

```
# effectifs : nombre de "F" et de "P"
effectifs <- table(x)
effectifs
```

```
## x
##  F  P
## 39 55
```

```
# diagramme en barre
barplot(effectifs,
        col = c("orange", "skyblue"),
        las = 1,
        ylab = "Effectifs",
        main = "Résultats du jeu de pile ou face")
```



On observe 55 piles et 39 faces sur 94 lancers. La proportion de piles (0,59) semble légèrement supérieure à celle attendue pour une pièce équilibrée (0,5).

Question 2

On suppose que le jeu n'est pas truqué, c'est-à-dire que la probabilité d'obtenir “F” est $p_0 = 1/2$. Pour chaque épreuve i, on définit la variable aléatoire X_i indiquant la réalisation de l'événement “obtenir F” :

$$X_i = \begin{cases} 1 & \text{si le résultat est F} \\ 0 & \text{sinon.} \end{cases}$$

Les épreuves sont supposées indépendantes. Justifier que la loi de la statistique

$$Z = \frac{1}{\sqrt{n}} \left(2 \sum_{i=1}^{n} X_i - n \right)$$

peut être approchée par la loi normale $\mathcal{N}(0,1)$.

Réponse.

On commence par calculer la moyenne et la variance de Z. On pose $S_n = \sum_{i=1}^n X_i$. On a

$$\mathbb{E}[Z] = \frac{2}{\sqrt{n}} \big(\mathbb{E}[S_n] - \tfrac{n}{2} \big) = 0\,,$$

et

$$\mathrm{Var}(Z) = \Big(\frac{2}{\sqrt{n}} \Big)^2 \mathrm{Var}(S_n) = \frac{4}{n} \times \frac{n}{4} = 1\,.$$

Justification de l'approximation normale :

Les conditions du TCL sont triviales ici (variance finie). La variable Z est une version centrée et réduite d'une somme de variables de Bernoulli identiques indépendantes de variance 1/4.

Nous avons

$$Z = \frac{S_n - n/2}{\sqrt{n/4}}.$$

Par le **théorème central limite**, la loi de Z est approchée par $\mathcal{N}(0,1)$.

Question 3

À partir des données observées, calculer la **valeur observée** de la statistique

$$z = \frac{1}{\sqrt{n}}\left(2\sum_{i=1}^{n} x_i - n\right),$$

ainsi que la valeur correspondante z^2. (On rappelle que $x_i = 1$ si le résultat est "F" et $x_i = 0$ sinon.)

Réponse. Inclure un code R pour calculer les valeurs demandées arrondies à 4 décimales et décrire brièvement les résultats obtenus.

```
n <- length(x)

# Calcul de z et z^2
z <- (1 / sqrt(n)) * (2 * sum(x == "F") - n)
z2 <- z^2

res <-c(z, z2)
names(res) <- c("z", "z-squared")
round(res,4)
```

```
##         z z-squared
##   -1.6503    2.7234
```

La valeur observée z est environ égale à -1,6503 et la valeur observée z^2 est environ égale à 2,7234.

Question 4

On suppose que la statistique Z suit la loi normale standard $\mathcal{N}(0,1)$. À partir de la valeur observée z obtenue à la Question 3, calculer numériquement à l'aide de R la probabilité

$$P = \mathbb{P}(|Z| > |z|).$$

Réponse. Utiliser la fonction `pnorm` pour calculer cette probabilité (arrondir à 4 décimales) et décrire brièvement le résultat obtenu.

```
# Probabilité bilatérale : P(|Z| > |z|)
P <- 2 * pnorm(abs(z), lower = FALSE)
round(P, 4)
```

```
## [1] 0.0989
```

La valeur de la probabilité P est environ égale à 0,0989.

Question 5

Interprétez la valeur de P obtenue à la Question 4 dans le cadre d'un test d'hypothèse sur la probabilité d'obtenir "Face". Précisez :

— l'hypothèse nulle et l'hypothèse alternative testées,
— la décision que vous prenez au seuil de 5 %.

Réponse.
La probabilité P correspond à la p-valeur du test d'hypothèse

$$H_0 : p = 0{,}5 \quad \text{(le jeu est équilibré)} \qquad H_1 : p \neq 0{,}5 \quad \text{(le jeu est truqué)}.$$

La p-valeur obtenue à la Question 4 est $P \approx 0{,}10$. Comme $P > 0{,}05$, on ne rejette pas l'hypothèse nulle au seuil de 5 %. Les données ne permettent donc pas de conclure que le jeu est truqué.

Question 6

Calculer la p-valeur du test de l'hypothèse nulle "le jeu est équilibré" contre l'hypothèse alternative "le jeu est truqué" à l'aide de la commande R `prop.test`. Comparer le résultat à celui obtenu précédemment et préciser quelle loi statistique est utilisée pour le calcul du test.

Réponse. Compléter et exécuter le code suivant.

```r
# Modifier eval = FALSE en eval = TRUE
# Test de proportion sans correction de continuité
test_trucage <- prop.test(x = 39,     # nombre de "F"
                          n = 94,     # nombre total d'épreuves
                          p = 0.5,    # hypothèse nulle : jeu équilibré
                          correct = FALSE)
test_trucage
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  39 out of 94, null probability 0.5
## X-squared = 2.7234, df = 1, p-value = 0.09889
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.3205503 0.5159198
## sample estimates:
##         p
## 0.4148936
```

La p-valeur est identique à la valeur P calculée dans la Question 5. Cela suggère que `prop.test` applique une approximation gaussienne pour calculer la p-valeur.

Question 7

À quoi correspond la statistique `X-squared = 2.7234` affichée dans la sortie de la commande prop.test ? Le test effectué à la Question 6 est-il équivalent au test réalisé à la Question 5 ?

Réponse. Justifiez brièvement votre réponse à l'aide d'un argument de probabilité.

La statistique `X-squared = 2.7234` correspond à la valeur z^2 calculée dans la Question 3. Ce test est équivalent à celui réalisé à la Question 5 car

$$P = \mathbb{P}(|Z| > |z|) = \mathbb{P}(Z^2 > z^2)\,.$$

En conclusion, le test de proportion `prop.test` repose sur la même approximation normale, simplement exprimée sous forme de Z^2 (ou χ^2) plutôt que de Z.

Question 8

Déterminer, à l'aide de R, un intervalle de confiance à 95 % pour la fréquence de "Face" ("F") observée dans les données (arrondir à 4 décimales). Indiquer la commande utilisée et interpréter brièvement le résultat.

Réponse.

On extrait `conf.int`de l'objet créé par la commande précédente.

```
# Intervalle de niveau de confiance 95%
round(test_trucage$conf.int, 4)
```

```
## [1] 0.3206 0.5159
## attr(,"conf.level")
## [1] 0.95
```

On remarque que la valeur 0,5 se trouve dans la plage des valeurs compatibles avec l'hypothèse nulle. On ne rejette pas l'hypothèse que le jeu est équilibré.

Question 9

Sous l'hypothèse d'un jeu équilibré, calculer les effectifs attendus de "F" et de "P" pour $n = 94$ observations. On note E_F et E_P les effectifs attendus et O_F et O_P les effectifs observés. Calculer la statistique du Chi-deux

$$\chi^2 = \frac{(E_F - O_F)^2}{E_F} + \frac{(E_P - O_P)^2}{E_P}\,.$$

À l'aide d'une commande R, effectuer un test du Chi-deux et calculer la p-valeur. Commenter le résultat obtenu.

Réponse.

```
# Modifier eval = FALSE en eval = TRUE
# effectifs attendus
n <- length(x)
E <- c(F = n * 0.5, P = n * 0.5)

# effectifs observés
O <- table(x)

# statistique du Chi-deux
chi2 <- sum((E - O)^2 / E)
chi2
```

```
## [1] 2.723404
```

La statistique du Chi-deux est égale à 2,7234. On retrouve la valeur calculée dans les questions précédentes (z^2 ou `X-squared`).

```
# Test du Chi-deux
chisq.test(table(x), p = c(0.5, 0.5))
```

```
##
##  Chi-squared test for given probabilities
##
## data:  table(x)
## X-squared = 2.7234, df = 1, p-value = 0.09889
```

Le test est equivalent à celui effectué Question 7. Il repose sur une approximation gaussienne. Au seuil de 5 %, les données ne fournissent pas de preuve suffisante que le jeu soit truqué.

Question 10

On révèle que le jeu de données provient en réalité de la simulation suivante :

```r
set.seed(2025)
x_sim <- sample(c("F", "P"), 94, prob = c(0.40, 0.60), replace = TRUE)

# verification
identical(x, x_sim)
```

```
## [1] TRUE
```

Sachant que la probabilité réelle d'obtenir "F" est $p = 0,40$, que peut-on conclure sur les valeurs de significativité (p-valeurs) obtenues dans les Questions précédentes ? Commentez brièvement la validité de l'analyse statistique.

Réponse. Les tests précédents reposaient sur l'hypothèse nulle ($H_0 : p = 0{,}5$). Les p-valeurs obtenues (10 %) étaient supérieures au seuil de 5 %, ce qui nous a conduit à ne pas rejeter (H_0). En réalité, les données proviennent d'une loi où $p = 0,4$: le jeu est donc effectivement légèrement biaisé en faveur de "P". Cela montre que, malgré un biais réel, la taille de l'échantillon ($n = 94$) est trop faible pour détecter ce biais au seuil de 5 %. On a commis ici une erreur de type "ne pas rejeter une hypothèse fausse".