

Corrigé examen atelier de statistiques Cogmaster

Tous documents autorisés. Durée de l'épreuve = 2h

1 Questions

1. La moyenne d'un échantillon de 10 nombres distribués selon une loi normale centrée réduite est exactement zéro.

oui non

2. La p -value calculée dans un test d'hypothèse représente :

la probabilité que l'effet observé corresponde à un effet réel dans la population.

la probabilité que la statistique de test observée soit aussi extrême sous l'hypothèse nulle.

1 - la probabilité de l'hypothèse alternative.

3. Dans un test d'hypothèse, faire une « fausse alarme », c'est :

rejeter l'hypothèse nulle alors qu'elle est vraie.

accepter l'hypothèse nulle alors qu'elle est fausse.

accepter l'hypothèse alternative alors qu'elle est vraie

rejeter l'hypothèse alternative alors qu'elle est fausse

4. Le test T de Student peut être utilisé pour déterminer si la moyenne d'un ensemble de mesures continues suivant à peu près une distribution normale est différente de zéro.

oui non

5. Si, à un examen comprenant 10 questions à choix binaires, un étudiant a répondu absolument au hasard, quel est le nombre N de réponses correctes le plus probable ? A l'aide de R, calculer la probabilité d'observer précisément ce nombre de réponses.

$N= 5$ $\text{prob}(n=N / \text{étudiant répondant au hasard})= 0.25$ (`dbinom(5,10,1/2)`)

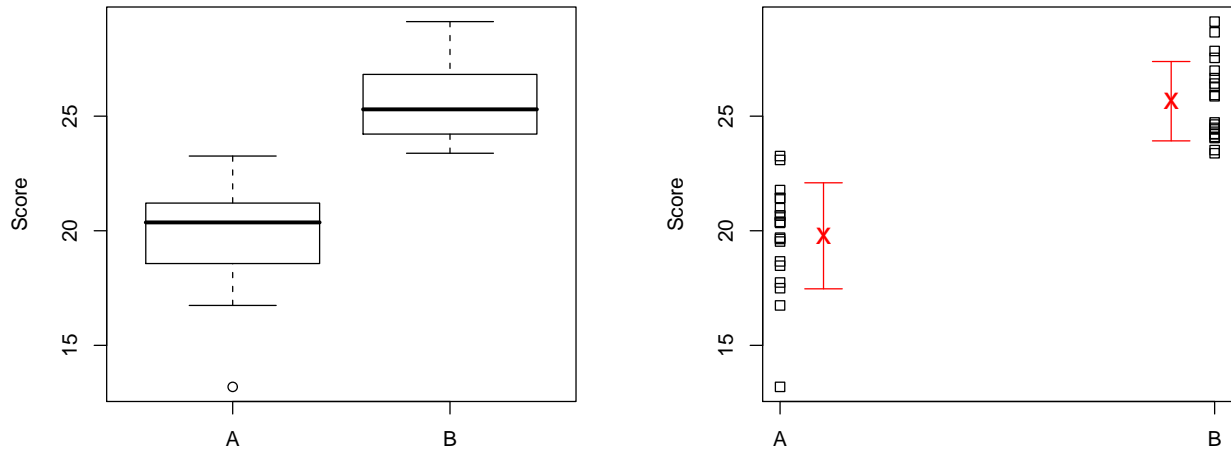
6. Quelle est la probabilité qu'un étudiant répondant totalement au hasard obtienne exactement 10/10 réponses correctes ?

$\text{prob}(n=10 / \text{étudiant répondant au hasard})= 1/1000$ (`dbinom(10,10,prob=1/2)`)

7. Mêmes questions que les deux précédentes pour un étudiant qui, ayant travaillé, avait *a priori* 2/3 de chances de répondre correctement à chaque question.

$N= 7$ $\text{prob}(N / \text{étudiant répondant avec } 2/3) = 0.26$
($N=7$ obtenu par `barplot(dbinom(0:10,10,2/3))`)

8. Les résultats ci-dessous représentent des scores observés entre deux groupes indépendants de sujets. Quel est le score médian observé pour le groupe A ? Quel est (à peu près) le score maximum observé dans le groupe B ? Le graphique de droite est un graphique des données brutes et des moyennes \pm écart-type. Un test t de Student est-il *a priori* justifié dans ce contexte ? Si oui, à votre avis le test serait-il significatif ?



Réponse :

Le score médian du groupe A est légèrement supérieur à 20.

Le score maximum du groupe B est à peu près 30.

On peut appliquer un test de Student car les distributions paraissent normales et de variances semblables (évidemment, pas moyen de vérifier l'hypothèse d'indépendance).

Vu que les distributions ne se recouvrent pratiquement pas, on s'attend à ce que la différence soit significative. Plus précisément, la taille de l'effet (différence entre les deux moyennes) est de l'ordre de 2 écart-types, donc la valeur de T qui est égale à la différence des moyennes divisée par l'écart-type et par \sqrt{N} (où N est le nombre de mesures), est donc très supérieure à deux, donc certainement significatif (le seuil bilatéral à 5% pour un T à 10 degrés de liberté est de 2.3, obtenu par `qt(0.975, 10)`).

2 Exercices

(Ecrivez les commandes R utilisées et (de manière succincte) les résultats produits)

2.1 Exercice 1

Récupérer les données du fichier `www.pallier.org/stats.2009/interro/exam.dat`. Il fournit les résultats (réussite ou non) d'un certain nombre d'élèves à un examen (1 ligne par élève).

Combien y-a-t-il d'élèves en tout ?

```
> a <- read.table("exam.dat", header=T)
> nrow(a) ou length(a$eleve)
[1] 233
```

Quel est le pourcentage de réussite global ?

```
> mean(a$reussi=='oui')
[1] 0.4034335
```

Combien y-a-t-il de garçons et de filles ?

```
> table(a$genre)
  fille garçon
  132    101
```

Calculer les pourcentages de réussite en fonction du genre (garçon ou fille) ?

```
> with(a, tapply(reussi=="oui", genre, mean))
  fille    garçon
0.3787879 0.4356436
```

L'effet du genre sur la réussite est-il significatif au seuil de 5% ?

```
> prop.test(table(a$genre, a$reussi))
```

2-sample test for equality of proportions with continuity correction

```
data: table(a$genre, a$reussi)
X-squared = 0.5504, df = 1, p-value = 0.4581
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.07915777  0.19286914
sample estimates:
 prop 1    prop 2
0.6212121 0.5643564
```

La différence n'est pas significative à 5% ($p\text{-val}=0.46$).

2.2 Exercice 2

Récupérer le tableau de données `www.pallier.org/stats.2009/interro/scores.dat`

– Combien y-a-t-il de sujets différents, de groupes différents, de scores par sujet ?

```
> a<-read.table("scores.dat",header=T)
> attach(a)
> length(unique(suj))
[1] 50
```

```
> length(unique(group))
[1] 5
> table(suj)
suj
 suj1 suj10 suj11 suj12 suj13 suj14 suj15 suj16 suj17 suj18 suj19  suj2 suj20
   10   10   10   10   10   10   10   10   10   10   10   10   10
...
```

– Le facteur « suj » est-il emboité ou croisé avec le facteur « group » ?

`table(suj, group)` permet de constater que suj est emboité dans group

– Construire un `data.frame` avec une ligne par sujet et 3 colonnes : sujet, groupe et la médiane des scores de ce sujet.

```
> dd <- aggregate(score, list(suj=suj, group=group), median)
```

– Effectuer une Anova pour comparer les groupes (fournir la valeur de F et la p-value).

```
> summary(av <- ov(x ~ group, data=dd))
              Df Sum Sq Mean Sq F value    Pr(>F)
group           4 790.65  197.66  49.793 5.766e-16 ***
Residuals      45 178.64    3.97
```

L'effet de groupe est significatif ($F(4,45)=49.8$; $p<0.001$)

– Afficher les intervalles de confiance de Tukey.

```
plot(TukeyHSD(av))
```

2.3 Exercice 3

Récupérer le tableau de données www.pallier.org/stats.2009/interro/Davis.dat.

Ce tableau contient les variables suivantes pour 200 personnes (attention : il y a des données manquantes)

sex A factor with levels : 'F', female; 'M', male.

weight Measured weight in kg.

height Measured height in cm.

repwt Reported weight in kg.

repht Reported height in cm.

– Résumer numériquement et graphiquement les données (Au moins un histogramme par variable numérique, et des boîtes à moustaches pour chaque variable numérique en fonction de la variable catégorielle (sexe)). Isoler la ou les observation(s) aberrante(s) pour chacune des variables numériques, i.e. donner le ou les numéro(s) d'observation correspondant(s). Quel est le sexe des individus n'ayant pas indiqué de poids/taille perçus (faire un tableau d'effectifs) ?

```
> a<-read.table('Davis.dat',header=T)
> attach(a)
```

```

> par(mfcol=c(2,2))
> hist(weight); hist(height); hist(repwt); hist(repht)
> boxplot(weight~sex); boxplot(height~sex)
> boxplot(repwt~sex); boxplot(repht~sex)

```

Données aberrantes. Sur les plots suivants :

```
plot(weight~repwt); plot(height~repht)
```

on identifie deux points aberrants ($\max(\text{weight})$ et $\min(\text{height})$) : On peut utiliser `identify()` sur la fenetre graphique, ou :

```

> which(weight==max(weight))
[1] 12
> which(height==min(height))
[1] 12
> a[12,]
   sex weight height repwt repht
12  F    166     57    56    163

```

On constate que les poids et tailles ont été échangés pour cette personne. On peut soit corriger l'erreur, soit supprimer la ligne (voir question suivante).

- Après avoir supprimé les données manquantes et la ou les valeur(s) aberrante(s) détectée(s) en 1, construire un modèle de régression linéaire poids perçu \sim poids réel. Commenter les coefficients de régression. Construire deux modèles séparés : un pour les hommes et un pour les femmes.

```

> a <- a[-12,]
> summary(a)
> b <- a[!(is.na(a$repwt)|is.na(a$repht)),]
> plot(b$repwt~b$weight)
> summary(lm1 <- lm(repwt ~ weight, data=b))
> abline(lm1)
> abline(0,1, lty=2)

```

Les coefficients ne sont pas significativement différents de 0 (intercept) et 1 (pente). On ne peut pas dire que les gens sur-estiment ou sous-estiment systématiquement leur poids.

```

> summary(lm2 <- lm(repwt ~ weight, data=b, subset=sex=="M"))
> summary(lm3 <- lm(repwt ~ weight, data=b, subset=sex=="F"))

```

Là encore, pas de différence majeure entre hommes et femmes (on peut le tester formellement avec un modèle `lm(repwt~weight*sex, data=b)`, mais cette analyse de covariance n'a pas été vue en cours)

- Calculer un score de différence taille perçue–taille réelle et comparer ces scores entre hommes et femmes. La différence entre les scores moyens est-elle significative au seuil 5 % ?

```

> x=b$repht-b$height
> tapply(x, b$sex, summary)
$F
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-10.000 -3.000  -3.000  -2.378  -1.250   6.000

$M
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  -8.000 -3.000  -2.000  -1.756   0.000   3.000
> boxplot(x~b$sex)
> stripchart(x~b$sex, method='jitter')
> t.test(x~b$sex)

```

Welch Two Sample t-test

data: x by b\$sex

t = -2.0008, df = 167.087, p-value = 0.04703

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-1.23465125 -0.00825567

sample estimates:

mean in group F mean in group M

-2.377551 -1.756098

Les femmes sous-estiment leur taille plus que les hommes.