

Notes sur les Statistiques

Paviel Schertzer

Ing1 2022

Notes de rendu de l'examen

Examen (DM de 1 heure 30)

Le rendu

Séparateur des décimales avec une virgules.

Pour les réponses en **OUI** et **NON**

Ne pas mettre de friture.

Quand on aura formulé toutes les réponses, il va falloir le sauvegarder. Le nom du fichier est fixé par Epita.

On dépose, Epita fait les archives, puis le prof télécharge et fait les corrections. On met le nom et le prénom.

Le format est **csv** avec séparateur **;**. Donc le **Format français**.

Le récapitulatif sera fait dans l'énoncé. Ce qui est important maintenant est de faire le test avec la licence chez nous après le cours.

FORMAT : CSV Séparateurs ;

Nous avons une licence Microsoft à Epita.

==connaître son numéro UID==

Pas deux personnes auront le même numéro d'application ni le même résultat.

Être capable de jouer avec R

On doit pouvoir jouer avec R. On a tous les supports de R.

Lancer une commande et lire le résultat.

4 choses à savoir faire :

- ☐ jouer avec R
- ☐ savoir ce qu'est un intervalle de confiance
- ☐ savoir ce qu'est un test
- ☐ savoir ce qu'est une régression linéaire

Date de rendu

Le professeur est d'accord pour favoriser Juillet.

Cependant Epita a un planning très chargé, et il faut vérifier dans la semaine que l'on a tout ce qu'il a dit.

Liens utiles

Voir le live

<https://www.twitch.tv/herbertgroscot>

Regarder les rediffusions

<https://web.microsoftstream.com/user/2ff366cc-a790-4ccc-be30-83b757debb2a>

Fichiers (entretenu par les délégués)

https://epitafr-my.sharepoint.com/personal/florian_cecilon_epita_fr/_layouts/15/onedrive.aspx?id=%2Fpersonal%2Fflorian%5Fcecilon%5Fepita%5Ffr%2FDocuments%2FEpita%20%2D%202022%2FCours%2FStatistiques&originalPath=aHR0cHM6Ly9lcGl0YWZyLW15LnNoYXJlcG9pbnQuY29tLzpmOi9nL3BlcnNvbmFsL2Zsb3JpYW5fY2VjaWxvbl9lcGl0YV9mci9FbFlwRmh2WDE3SkRwSEg3ZkhtaC1YZ0JDZGpNcThmSmhPUmQxb2hOSGhnQkFnP3J0aW1IPTFtb2E4UDBHMkVn

Live Coding

Exo1

```
> UID<-20254
> X<-runif(1000)
> Z<-1:1000
> plot(Z)

> alpha->UID/23000
> K<-UID/75000

> alpha[1]
> ## à compléter
```

Exo 2

```
> ## poid d'un nouveau né
> ## on en a pesé 49
> ## on a trouvé une moyenne de 6,6 kg
>
> ## "Je sais que l'écart-type est de 0,5 kg"
>
> ## Intervalle de confiance à 95% du poids moyen d'un bébé
>
> ## Une dernière hypothèse : le poid suit une loi normale !
>
> ## votre camarade a répondu  $m - 2\sigma$ ,  $m + 2\sigma$ 
```

```

> ## C'est pasiditio dans la mesure où l'on mesure la moyenne. Quans on fait
une expérience on connaît l'espérance et on observe une variation de
l'espérance due à l'écart type.
>
> ## en général "Observation = moyenne + écart = moyenne + k * exart type"
>
> ## estimation de moyenne de type moyenne observée +- k * ecart-type
> ## Quand on connaît l'écart-type : k dépend de la distribution de la loi
normale.
>
>
> ## Au lieu de dire je connais l'espérance et je fais une simulation, je vais
dire j'ai mesuré une moyenne à partir des observations, et la vraie moyenne
d'écarte de la moyenne avec une valeur par rapport à l'écart type.
>
> ## Pour cela je vais prendre le ours "Le contrôle..." -> "Estimation "
>
> x.barre<-3.6
> sigma0<-0.5
> n<-49
>
> ## Un intervalle de confiance à 95%
> ## Le taux d'erreur (cf convention des taux d'erreurs)
> ## alpha=5%
> ## alpha/2=2.5%
> ## 1-alpha/2 = 0.975
>
> ## La fonction R qui permet de le faire est qnorm
>
> qnorm(0.975)
[1] 1.959964
>
> u<-qnorm(0,975)
>
> u
[1] -Inf
> u<-qnorm(0.975)
> u
[1] 1.959964
>
> ## reste à appliquer la slide du cours sur l'intervalle de confinace
> mu.inferieur<-x.barre-u*sigma0/sqrt(n)
> mu.inferieur
[1] 3.460003
> mu.superieur<-x.barre+u*sigma0/sqrt(n)
> mu.superieur
[1] 3.739997
> # Intervalle de confiance à 95%
> ## borne inférieure à 3.46 et supérieure à 3,47

```

```

>
> ## -----<Question Complementary>-----
> ## Changement de sujet
> ## On a mesuré un écart type de 0,53 kg
> ## Quel est l'intervalle de confiance ? (Dire ce qu'il change, ça a été fait
en cours)
>
> ## Quand la variance est connue on a une variance qui suit une loi normale.
Si elle est inconnue elle ne suit plus la loi normale. C'est toujours une
courbe en cloche mais quand on a peu d'individus dans l'échantillon, on a une
incertitude en raison du nombre, mais aussi sur... et elle existe par
programme, et elle s'appelle QT.
> ## On aura la même chose mais un peu différente, et au lieu de diviser par
racine de n, on va diviser par la racine de n-1.
> v<-qt(0.975,49-1)
> v
[1] 2.010635
> ecart<- v * 0.53/sqrt(n - 1)
> mu.inferieur<-x.barre - ecart
> mu.inferieur
[1] 3.446189
> mu.superieur<-x.barre + ecart
> mu.superieur
[1] 3.753811

```

Exo 3

```

> ## Petit test
> ## Pour une maladie donnée, on me dit qu'un traitement guerit 90% des
patients
> ## (ps : dans la vraie vie c'est plus compliqué, on a l'état du patient,
l'âge, ...)
> ## J'ai fait un test avec 1000 patients
> ## Il y en 850 de guéris
>
> ## (au bout de deux semaines)
>
> ## J'accepte le 90% sur cette base ?
> ## Test de chi2, qui est dans le cours
>
> ## échantillons de n individus rangés dans k classes
> ## ici : k=2 classes : 1 : patients guéris, 2. non guéris
> ## on m'a dit          p1 = 0.9          p2 = 0.1
>
> ## J'ai observé          850          150
>
> n<-1000
> N1<-850

```

```

> N1<-150
> p1<-0.9
> p2<-0.1
>
> Z<-(N1-n**p1)^2/(n*p1) + (N2-n*p2)^2/(n*p2)
Erreur : unexpected input in "Z<-(N1-n**"
> Z<-(N1-n*p1)^2/(n*p1) + (N2-n*p2)^2/(n*p2)
Erreur : objet 'N2' introuvable
> N1<-850
> N2<-150
> Z<-(N1-n*p1)^2/(n*p1) + (N2-n*p2)^2/(n*p2)
> Z
[1] 27.77778
> qchisq(0.95, 1)
[1] 3.841459
> qchisq(0.99, 1)
[1] 6.634897
> pchisq(27.777, 1)
[1] 0.99999999
> 1-pchisq(27.777, 1)
[1] 1.361349e-07
> ## Pour rester dans le cours, on va dire que la valeur de Z, des écart est
trop grande, et que l'on va refuser en principe
> ## En principe Z doit rester petit
> ## Ici "Z est trop grand"
> ## On refuse l'hypothèse
>
> ## On refuse l'hypothèse de guérison à 90%
>
> ## H0 : la proba de guérison est de 90%, la proba de non guérison est de 10%

```

Exo 4 : régression

Le prof a généré des valeurs, il est donc impossible de compiler dans la console sans les générer.

```

> ## exemple bidon avec 1000 variables réparties
> ## Les données nous seront données

```

```

> plot(X, Y, col = "blue")
> mX<-mean(X)
> mY<-mean(Y)
> sX<-sd(X)
> sY<-sd(Y)
> rho<-cor(X, Y)

```

```

> mX
[1] 489.5764
> mY
[1] 2450.988
> rho
[1] 0.9194288
> sX
[1] 279.0857
> xY
[1] 1474.709

```

En utilisant la slide **Modèle (2)**

$$\hat{\beta} = r \frac{s_y}{s_x}$$

$$\hat{a} = \bar{Y} - \hat{\beta} \bar{X}$$

```

> beta<-rho*sY/sX
> beta
[1] 4.858326

> alpha<--mx - beta*mX
> alpha
[1] 72.46558

> PREV<-beta*X+alpha
> points(X, PREV, col="red", pch=19, cex=0.8)
> # donne la droite de régression
> ECARTS<-Y - PREV

> var(Y)
[1] 2174766
> var(PREV)
[1] 1838437
> var(ECARTS)
[1] 336329
> 1838437+336329
[1] 2174766

> var(PREV)/var(Y)
[1] 0.8453493

```

Récapitulatif

- on aura un fichier pdf avec l'énoncé
- on rend un fichier `.csv` avec séparateur `;`
- on utilise `R`

- on respecte les conventions données par Epita