SYS865 Inférence statistique avec programmation R

Ornwipa Thamsuwan

14 février 2024

Recap et plan

Test sur la variance des deux échantillons

Test de normalité

Travaux pratiques

Recap et plan

Recap et plan

Les derniers cours ...

- Variables aléatoires
- Échantillonage
- Inférence statistique
  - Intervalle de confiance
  - Types d'erreur
  - Tests d'hypothèse
    - Test sur la moyenne d'un échantillon
    - Test sur la moyenne des deux échantillons
    - Test nonparamétrique
  - Valeur p

Dans ce cours ...

- Tests pour les conditions des statistiques paramétriques
  - Test d'hypothèse sur la variance des deux échantillons
  - Test de normalité
- Accompagnement du projet

Test sur la variance des deux échantillons

Test sur la variance des deux échantillons

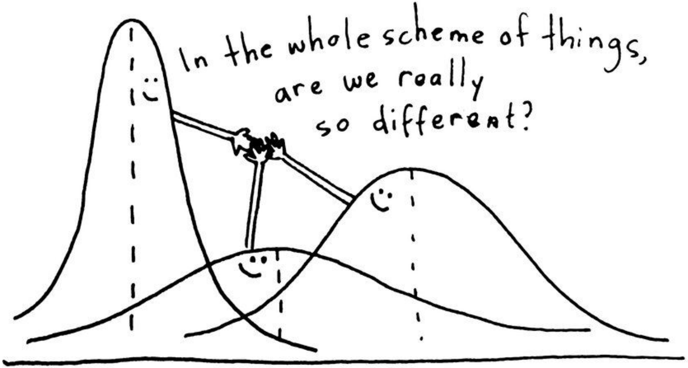


Figure 1: Homogénéité de la variance

Test F

Contexte statistique: Le test F est utilisé pour comparer les variances de deux échantillons indépendants afin de déterminer si elles sont significativement différentes. Il est souvent utilisé dans le contexte d'une ANOVA, mais peut également être utilisé seul.

Test F

Contexte statistique: Le test F est utilisé pour comparer les variances de deux échantillons indépendants afin de déterminer si elles sont significativement différentes. Il est souvent utilisé dans le contexte d'une ANOVA, mais peut également être utilisé seul.

Équation mathématique: La statistique de test pour un test F est calculée comme suit :

$$F = \frac{Var(X_1)}{Var(X_2)}$$

Où :

- $Var(X_1)$ et $Var(X_2)$ sont les variances des échantillons des deux échantillons indépendants.
- F est la statistique de test qui suit une distribution F sous l'hypothèse nulle.

Recap et plan

Test sur la variance des deux échantillons

Test de normalité

Travaux pratiques

Distribution F

La forme de la distribution F dépend des degrés de liberté df_1 et df_2.

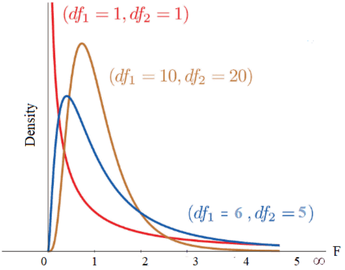


Figure 2: Distribution F

Les degrés de liberté pour le numérateur sont $df_1 = n_1 - 1$ et pour le dénominateur $df_2 = n_2 - 1$, où n_1 et n_2 sont les tailles des échantillons des deux échantillons.

Test F

Hypothèse nulle

L'hypothèse nulle affirme que les deux variances sont égales. Mathématiquement, elle est exprimée comme suit :

$$H_0 : \sigma_1^2 = \sigma_2^2$$

Où σ_1^2 et σ_2^2 sont les variances des deux populations.

Test F

Hypothèse nulle

L'hypothèse nulle affirme que les deux variances sont égales. Mathématiquement, elle est exprimée comme suit :

$$H_0 : \sigma_1^2 = \sigma_2^2$$

Où σ_1^2 et σ_2^2 sont les variances des deux populations.

Hypothèse alternative

L'hypothèse alternative peut être bilatérale ou unilatérale :

$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

$$H_1 : \sigma_1^2 < \sigma_2^2$$

$$H_1 : \sigma_1^2 > \sigma_2^2$$

Règle de décision : Afin de déterminer si les variances sont significativement différentes, on compare la valeur F calculée à la valeur critique de la table de distribution F à un certain niveau de signification (α, souvent 0,05).

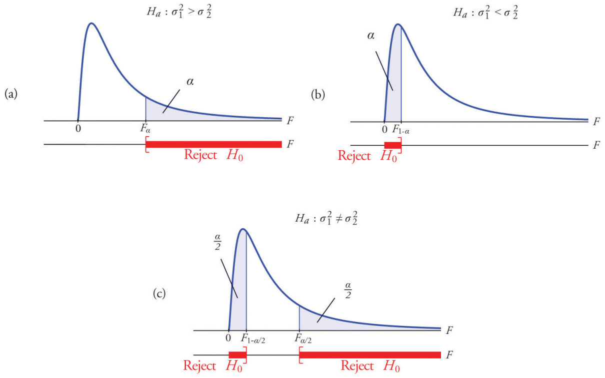


Figure 3: Rejeter H_0

Analyse avec R

Considérant le paramètre `Glucose` sur la base de données "Pima Indian Diabetes", les variances des deux groupes de `Outcome` sont-elles égales ?

Analyse avec R

Considérant le paramètre `Glucose` sur la base de données "Pima Indian Diabetes", les variances des deux groupes de `Outcome` sont-elles égales ?

Compter la taille de l'échantillon pour chaque groupe.

```r
data <- read.csv("diabetes.csv")
filtered_data <- subset(data, Glucose > 0)
table(filtered_data$Outcome)
```

```
##
##   0   1
## 497 266
```

Analyse avec R

Considérant le paramètre `Glucose` sur la base de données "Pima Indian Diabetes", les variances des deux groupes de `Outcome` sont-elles égales ?

Compter la taille de l'échantillon pour chaque groupe.

```
data <- read.csv("diabetes.csv")
filtered_data <- subset(data, Glucose > 0)
table(filtered_data$Outcome)
```

```
##
##   0   1
## 497 266
```

Le degré de liberté $df = n - 1$

```
outcome_counts <- table(filtered_data$Outcome)
df0 <- outcome_counts["0"] - 1
df1 <- outcome_counts["1"] - 1
```

Analyse avec R

Calculer la variance pour chaque groupe.

```
variances <- tapply(filtered_data$Glucose,
                    filtered_data$Outcome, var)
variances
```

```
##        0        1
## 613.8951 876.1126
```

Analyse avec R

Calculer la variance pour chaque groupe.

```
variances <- tapply(filtered_data$Glucose,
                    filtered_data$Outcome, var)
variances
```

```
##        0        1
## 613.8951 876.1126
```

La statistique de test $F = \frac{Var(X_1)}{Var(X_2)}$

```
F_statistics <- variances[1] / variances[2]
F_statistics
```

```
##         0
## 0.7007034
```

Analyse avec R

En case de “two-tailed” $H_1 : \sigma_1^2 \neq \sigma_2^2 \ldots$

Déterminer la valeur critique pour le test F à $\alpha = 0{,}05$.

```
alpha <- 0.05
lower_critical_value <- 1 / qf(1-alpha/2, df0, df1)
upper_critical_value <- qf(1-alpha/2, df0, df1)
```

```
cat(sprintf("Lower CV: %.3f, Upper CV: %.3f",
            lower_critical_value, upper_critical_value))
```

```
## Lower CV: 0.807, Upper CV: 1.240
```

Analyse avec R

En case de “two-tailed” $H_1 : \sigma_1^2 \neq \sigma_2^2 \ldots$

Déterminer la valeur critique pour le test F à $\alpha = 0{,}05$.

```
alpha <- 0.05
lower_critical_value <- 1 / qf(1-alpha/2, df0, df1)
upper_critical_value <- qf(1-alpha/2, df0, df1)
```

```
cat(sprintf("Lower CV: %.3f, Upper CV: %.3f",
            lower_critical_value, upper_critical_value))
```

```
## Lower CV: 0.807, Upper CV: 1.240
```

- `F_statistics` est inferieur à valeur critique inférieure.
- Rejeter H_0 et donc conclure que les variances des deux groupes ne sont pas égales.

Analyse avec R

Ou avec la fonctionne R: `var.test(group0, group1)`

```
##
##  F test to compare two variances
##
## data:  group0 and group1
## F = 0.7007, num df = 496, denom df = 265, p-value = 0.0007
## alternative hypothesis: true ratio of variances is not equ
## 95 percent confidence interval:
##  0.5653098 0.8625836
## sample estimates:
## ratio of variances
##          0.7007034
```

Puis observer la valeur p: `p-value < 0.05`

Test F

Considérations :

- Le test F suppose que les données des deux échantillons sont **normalement distribuées**.
- Les observations de chaque échantillon doivent être **indépendantes** les unes des autres. Une violation de cette hypothèse, comme cela peut se produire dans la conception appariée.

Test F

Considérations :

- Le test F suppose que les données des deux échantillons sont **normalement distribuées**.
- Les observations de chaque échantillon doivent être **indépendantes** les unes des autres. Une violation de cette hypothèse, comme cela peut se produire dans la conception appariée.

... Quoi faire quand des données ne sont pas normalement distribuées ?

Test de Levene

Le test de Levene évalue les différences entre les moyennes des écarts absolus des groupes par rapport à leurs moyennes ou médianes.

Hypothèse nulle $H_0 : \sigma_1^2 = \sigma_2^2$

Hypothèse alternative $H_1 : \sigma_1^2 \neq \sigma_2^2$

Test de Levene

Le test de Levene évalue les différences entre les moyennes des écarts absolus des groupes par rapport à leurs moyennes ou médianes.

Hypothèse nulle $H_0 : \sigma_1^2 = \sigma_2^2$

Hypothèse alternative $H_1 : \sigma_1^2 \neq \sigma_2^2$

La statistique suit une distribution F avec 1 et $N - 2$ degrés de liberté, où N est le nombre total d'observations.

Test de Levene

Le test de Levene évalue les différences entre les moyennes des écarts absolus des groupes par rapport à leurs moyennes ou médianes.

Hypothèse nulle $H_0 : \sigma_1^2 = \sigma_2^2$

Hypothèse alternative $H_1 : \sigma_1^2 \neq \sigma_2^2$

La statistique suit une distribution F avec 1 et $N - 2$ degrés de liberté, où N est le nombre total d'observations.

Une `p-value` inférieure à un seuil (généralement 0,05) indique des différences significatives dans les variances entre les groupes.

Analyse avec R

Le langage R a une bibliothèque pour le test de Levene.

```
library(car)
```

```
## Loading required package: carData
```

```
leveneTest(Glucose ~ factor(Outcome), data = filtered_data)
```

```
## Levene's Test for Homogeneity of Variance (center = median
##        Df F value    Pr(>F)
## group   1  23.212 1.752e-06 ***
##       761
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

Test de normalité

Test de normalité

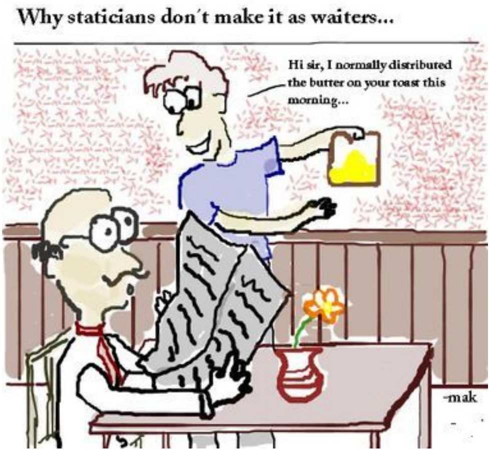


Figure 4: Normalité en statistique

Le moyen le plus rapide en langage R

Nous utilisons déjà le **Test de Shapiro-Wilk** dans les cours précédents.

```
shapiro.test(filtered_data$Glucose)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  filtered_data$Glucose
## W = 0.96964, p-value = 1.72e-11
```

Le moyen le plus rapide en langage R

Nous utilisons déjà le **Test de Shapiro-Wilk** dans les cours précédents.

```
shapiro.test(filtered_data$Glucose)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  filtered_data$Glucose
## W = 0.96964, p-value = 1.72e-11
```

H_0 est que les données sont normalement distribuées.

Une `p-value` faible (typiquement $< 0{,}05$) suggère que les données ne suivent pas une distribution normale.

Méthode graphique

Graphique Q-Q (Quantile-Quantile) montre les quantiles des données par rapport aux quantiles d'une distribution normale. Si les points se situent approximativement le long d'une ligne droite, cela suggère une normalité.

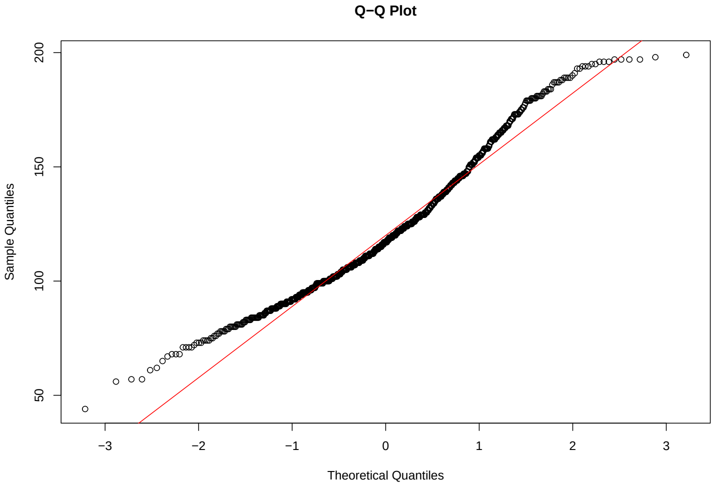
Méthode graphique

Graphique Q-Q (Quantile-Quantile) montre les quantiles des données par rapport aux quantiles d'une distribution normale. Si les points se situent approximativement le long d'une ligne droite, cela suggère une normalité.

- `qqnorm()` génère le graphique Q-Q, en traçant les quantiles de Glucose par rapport aux quantiles d'une distribution normale standard.
- `qqline()` ajoute une ligne de référence au graphique, ce qui facilite la visualisation des écarts par rapport à la normalité.

Graphique Q-Q

```r
qqnorm(filtered_data$Glucose, main = "Q-Q Plot")
qqline(filtered_data$Glucose, col = "red")
```



Travaux pratiques

Travaux pratiques

En divisant la base de données “Pima Indian Diabetes” en groupe de non diabétiques et diabétiques, pour chacun des huit paramètres (`Pregnancies`, `Glucose`, `BloodPressure`, `SkinThickness`, `Insulin`, `BMI`, `DiabetesPedigreeFunction` et `Age`) ...

1. Créer un graphique Q-Q pour tester la normalité des données
2. Tester l'homogénéité des variances en utilisant une méthode appropriée (soit le test F ou le test de Levene) en fonctionne de la normalité des données