Chapitre III. Tests de comparaison d'échantillons

Cours de Tests paramétriques

Deuxième Année - IUT STID - Olivier Bouaziz

2018-2019

Introduction

- Jusqu'à présent nous avons éffectué des tests à partir d'un seul échantillon. On comparait l'espérance, la variance ou la proportion d'un caractère d'une population à une valeur de référence.
- On souhaite à présent effectuer des tests de comparaison de deux populations différentes selon un critère quantitatif particulier à partir de deux échantillons extraits de ces deux populations. Par exemple :
  - Comparer le salaire moyen des femmes cadres et celui des hommes cadres.
  - Comparer les performances de deux machines au vu de la proportion de pièces défectueuses qu'elles produisent.
  - Comparer l'efficacité d'un traitement en comparant des mesures effectuées avant traitement à des mesures faites après traitement.
  - Comparer la durée de vie moyenne des PME en 2008 et en 1950.
- La généralisation à plus de deux populations sera faite dans le cadre du cours de l'analyse de la variance (modèle linéaire).

3.1 Test sur échantillons appariés

On parle de données appariées lorsque les deux échantillons considérés sont formés des individus, pour lesquels on a fait deux mesures d'une même quantité, généralement avec écart temporel et/ou après l'occurence d'un événement. On dispose donc de deux séries de données, de même taille $X_1, ..., X_{n_1}$ et $Y_1, ..., Y_{n_2}$, $(n_1 = n_2)$ où X_j et Y_j ont toutes deux été mesurées sur le $j-$ième membre de l'échantillon.

Exemple 1

Chez un échantillon de 30 sujets extrait d'une population, on mesure le rythme cardiaque (exprimé en pulsations par minute), noté X avant, et Y après administration d'un médicament, car on se demande si l'absorption de ce médicament a un effet sur le rythme cardiaque. On considère donc ici deux échantillons appariés (ou dépendants) $(X_1, \cdots, X_{30})$ et $(Y_1, \cdots, Y_{30})$. On obtient les résultats suivants :

patient	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
x	80	80	82	75	80	74	80	72	91	88	70	65	83	74	81
y	85	84	87	81	79	85	87	78	96	80	82	73	89	85	86
patient	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
x	68	69	71	70	73	78	75	76	78	77	75	72	71	75	78
y	72	74	77	75	81	70	77	76	82	83	80	80	81	76	77

L'absorption du médicament affecte-t-elle le ryhtme cardiaque de manière significative ?

- Ici, la modélisation est qu'on a d'une part $X_1, ..., X_n$ i.i.d distribuées selon une loi d'espérance μ_1 et d'autre part, $Y_1, \ldots, Y_n$ i.i.d. distribuées selon une loi éventuellemment différente d'espérance μ_2.
- Mais ici, on ne pourra supposer que les X_j sont indépendantes des Y_j (même sujet, j, qui influe à la fois sur X_j et Y_j).
- C'est pourquoi on s'intéressera plutôt aux différences :

$$Z_j = X_j - Y_j$$

- Les v.a. $Z_1, ..., Z_n$ sont i.i.d. selon une loi d'espérance $\mu = \mu_1 - \mu_2$.
- Il s'agit alors de tester $H_0 : \mu = 0$ contre : $H_1 : \mu \neq 0$ ou $H_1 : \mu > 0$ ou encore $H_1 : \mu < 0$, selon le contexte.
- Dans le cas des échantillons appariés on applique les techniques du chapitre précédent.

Application avec R

$> x = c(80, 80, 82, 75, 80, 74, 80, 72, 91, \cdots, 71, 75, 78)$
$> y = c(85, 84, 87, 81, 79, 85, 87, 78, 96, \cdots, 81, 76, 77)$
$> t.test(x, y, paired = T)$

Paired t-test

data : x and y

$$t = -5.327, \text{df} = 29, \text{p-value} = 1.022e - 05$$

alternative hypothesis : true difference in means is not equal to 0
95 percent confidence interval :

$$-6.319972 \quad -2.813362$$

sample estimates : mean of the differences -4.566667
Conclusion : L'absorption du médicament affecte le rythme cardiaque de manière extrêmement significative.

3.2 Test de comparaison de deux espérances sur échantillons indépendants

Exemple 2

Des essais cliniques sont menés auprès de 137 patients atteints d'une maladie pulmonaire sans gravité afin de tester l'efficacité d'un traitement. Le protocole est le suivant :

- Des exercices respiratoires sont prescrits à 67 patients choisis au hasard ainsi qu'un placebo (goupe témoin (A)).
- Les mêmes exercices respiratoires sont prescrits aux 70 autres patients ainsi que le traitement (groupe traité (B)).
- Au bout de trois mois, l'amélioration de la capacité pulmonaire de chaque patient est mesurée. L'amélioration est mesurée (voir tableau suivant) sur une échelle de 0 (pas d'amélioration) à 10 (récupération totale).

Exemple 2 (suite)

Amélioration	Groupe témoin (A)	Groupe traité (B)
0	2	0
1	8	0
2	4	3
3	7	0
4	14	10
5	9	14
6	5	13
7	4	17
8	7	10
9	2	3
10	5	0
Ensemble	$n_1 = 67$	$n_2 = 70$
	$\bar{x} = 4,776$	$\bar{y} = 5,671$
	$s_1^2 = 7,479$	$s_2^2 = 2,601$

Exemple 2 (suite)

- L'amélioration moyenne du groupe traité est supérieure de 0,9 points à celle du groupe témoin. Le problème est de savoir si cette différence moyenne d'amélioration entre les deux échantillons doit être attribuée aux bienfaits du traitement ou aux fluctuations d'échantillonnage (le même protocole sur d'autres individus n'aurait sans doute pas donné les mêmes résultats).
- Autrement dit, la différence observée entre les deux moyennes empiriques est-elle statistiquement significative ?
- Le caractère étudié est l'amélioration :
  - Soit X la v.a. qui associe à la personne choisie son amélioration sans traitement. L'espérance μ_1 et l'écart-type σ_1 de X sont inconnus.
  - Soit Y la v.a. qui associe à la personne choisie son amélioration après traitement. L'espérance μ_2 et l'écart-type σ_2 de Y sont inconnus.
- Il s'agit de construire un test de comparaison des deux espérances μ_1 et μ_2.

Exemple 2 (suite)

- Le laboratoire qui veut mettre sur le marché le traitement tient à l'hypothèse selon laquelle son médicament (groupe B) est plus efficace qu'un placebo (groupe A). L'intérêt du laboratoire est de contrôler l'erreur qui consiste à décider à tort que le placebo est plus efficace que le médicament et donc ainsi de poser le test sous la forme : $H_0 : \mu_1 \leq \mu_2$ contre $H_1 : \mu_1 > \mu_2$, ce qui revient à poser le test :

$$H_0 : \mu_1 = \mu_2 \text{ contre } H_1 : \mu_1 > \mu_2.$$

- En revanche, l'agence de sécurité sanitaire qui délivre l'autorisation de mise sur le marché du traitement, veut contrôler la probabilité de décider à tort que le médicament est plus efficace que le placebo. Cette agence pose le test sous la forme : $H_0 : \mu_1 \geq \mu_2$ contre $H_1 : \mu_1 < \mu_2$, ce qui revient à poser le test :

$$H_0 : \mu_1 = \mu_2 \text{ contre } H_1 : \mu_1 < \mu_2.$$

(l'agence cherche à minimiser la probabilité de mettre sur le marché un nouveau médicament qui n'a pas prouvé son efficacité)

Cadre général

On dispose de mesures d'une même grandeur (salaire, taille, etc.) sur deux échantillons extraits indépendamment de deux populations différentes :

- Population 1 : Le caractère est noté par X. Soit $x_1, ..., x_{n_1}$ les réalisations d'un échantillon de v.a. $(X_1, \ldots, X_{n_1})$ de taille n_1, extrait de cette population et tels que $\mathbb{E}(X_i) = \mu_1$ et $\mathrm{Var}(X_i) = \sigma_1^2$; on note $\overline{X}$ la moyenne empirique et S_1^2 la variance empirique de cet échantillon càd : $\overline{X} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i$ et $S_1^2 = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(X_i - \overline{X})^2$.
- Population 2 : Le caractère est noté par Y. Soit $y_1, ..., y_{n_2}$ les réalisations d'un échantillon de v.a. $(Y_1, \cdots, Y_{n_2})$, de taille n_2, extrait de la population 2 avec $\mathbb{E}(Y_i) = \mu_2$, et $\mathrm{Var}(Y_i) = \sigma_2^2$; on note $\overline{Y}$ la moyenne empirique et S_2^2 la variance empirique de cet échantillon càd : $\overline{Y} = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i$ et $S_2^2 = \frac{1}{n_2-1}\sum_{i=1}^{n_2}(Y_i - \overline{Y})^2$.
- les variables X_i et Y_j sont supposées indépendantes (mesures réalisées sur des individus nécessairement différents).

Problème posé

A l'aide des deux échantillons on veut comparer ces deux populations. Celà revient à comparer les paramètres des lois de probabilités des X_i et Y_j. Pour un test de comparaison sur les espérances, on se demande si l'espérance de le grandeur considérée est la même dans les deux populations. On veut donc tester

$$H_0 \ : \ \mu_1 = \mu_2$$

- contre l'alternative bilatérale $H_1 \ : \ \mu_1 \neq \mu_2$,
- ou contre l'alternative unilatérale $H_1 \ : \ \mu_1 < \mu_2$,
- ou contre l'autre l'alternative unilatérale $H_1 \ : \ \mu_1 > \mu_2$.

Problème posé

- Supposons qu'on s'intéresse à la taille des hommes et des femmes d'un pays ; on note μ_1 l'espérance de la taille des femmes adultes et μ_2 l'espérance de la taille des hommes adultes. On peut alors chercher à tester si la taille moyenne des hommes est significativement supérieure à celle des femmes.
- On teste donc : $H_0 \ : \ \mu_1 = \mu_2$ contre $H_1 \ : \ \mu_1 < \mu_2$.
- Il est naturel de fonder les tests de H_0 sur l'écart $\overline{X} - \overline{Y}$ entre les moyennes observées des deux échantillons.
- Sous l'hypothèse H_0, la différence observée $\overline{X} - \overline{Y}$ doit avoir une espérance nulle puisque $\mathbb{E}(\overline{X}) - \mathbb{E}(\overline{Y}) = \mu_1 - \mu_2 = 0$.
- Il faut donc connaître la loi de $\overline{X} - \overline{Y}$ sous H_0.

Deux cas à distinguer

- **Cas des grands échantillons de loi quelconque**
- **Cas des petits échantillons gaussiens**.

3.2.1 Grands échantillons de loi quelconque

- Si les tailles d'échantillons n_1 et n_2 sont grandes, on sait d'après le TLC que, approximativement, on a

$$\overline{X} \approx \mathcal{N}\left(\mu_1, \frac{\sigma_1^2}{n_1}\right) \quad \text{et} \quad \overline{Y} \approx \mathcal{N}\left(\mu_2, \frac{\sigma_2^2}{n_2}\right).$$

- Comme les deux échantillons sont indépendants, $\overline{X}$ et $\overline{Y}$ sont indépendants et approximativement sous H_0, on a

$$\overline{X} - \overline{Y} \approx \mathcal{N}\left(0, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right).$$

- Soit encore, sous H_0 :

$$T = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \approx \mathcal{N}(0,1).$$

- **Mais** on ne peut utiliser directement la v.a. T comme statistique de test que lorsque les variances des deux populations sont connues.

3.2.1 Grands échantillons de loi quelconque

Remarque

Il faut que les tailles des **deux** échantillons n_1 **et** n_2 soient suffisamment grandes pour pouvoir appliquer le TCL pour $\overline{X}$ et $\overline{Y}$. En pratique on considérera qu'il faut avoir $n_1 > 30$ **et** $n_2 > 30$.

Nous allons distinguer les trois cas suivants :

- Cas où les variances σ_1^2 et σ_2^2 sont connues.
- Cas où les variances σ_1^2 et σ_2^2 sont inconnues.
- Cas où les variances σ_1^2 et σ_2^2 sont inconnues mais égales.

a. Cas où les variances σ_1^2 et σ_2^2 sont connues (grands échantillons)

- **Test de $H_0 : \mu_1 = \mu_2$ contre $H_1 : \mu_1 \neq \mu_2$**

La statistique de test est

$$T = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Sous H_0, $T \approx \mathcal{N}(0,1)$. Sous H_1, T a tendance à prendre desvaleurs plus petites ou plus grandes.

Règle de décision :

- si $|T| > c_\alpha$ on rejette H_0,
- si $|T| \leq c_\alpha$, on ne rejette pas H_0.

Le seuil c_α est tel que $\mathbb{P}_{H_0}(|T| > c_\alpha) = \alpha$, *i.e.* tel que $\mathbb{P}_{H_0}(T > c_\alpha) = \alpha/2$. Le seuil c_α est le quantile d'ordre $1 - \alpha/2$ de la loi $\mathcal{N}(0,1)$.

a. Cas où les variances σ_1^2 et σ_2^2 sont connues (grands échantillons)

- **Test de $H_0 : \mu_1 = \mu_2$ contre $H_1 : \mu_1 > \mu_2$**

Sous H_0, $T = \dfrac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \approx \mathcal{N}(0,1)$. Sous H_1, T a tendance a prendre des valeurs plus grandes.

Règle de décision :

- si $T > c_\alpha$ on rejette H_0,
- si $T \leq c_\alpha$, on ne rejette pas H_0,

où c_α est tel que $\mathbb{P}_{H_0}(T_{n_1,n_2} > c_\alpha) = \alpha$. Le seuil c_α est le quantile d'ordre $1 - \alpha$ de la loi $\mathcal{N}(0,1)$.

a. Cas où les variances σ_1^2 et σ_2^2 sont connues (grands échantillons)

- **Test de $H_0 : \mu_1 = \mu_2$ contre $H_1 : \mu_1 < \mu_2$**

Sous H_0, $T = \dfrac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \approx \mathcal{N}(0,1)$. Sous H_1, T a tendance a prendre des valeurs plus petites.

Règle de décision :

- si $T < c_\alpha$ $(c_\alpha < 0)$ on rejette H_0,
- si $T \geq c_\alpha$, on ne rejette pas H_0,

où c_α est tel que $\mathbb{P}_{H_0}(T < c_\alpha) = \alpha$. Le seuil c_α est le quantile d'ordre α de la loi $\mathcal{N}(0,1)$.

b. Cas où les variances σ_1^2 et σ_2^2 sont inconnues (grands échantillons)

- On remplace respectivement σ_1^2 et σ_2^2 par leurs estimateurs consistants et sans biais

$$S_1^2 = S_X^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_i - \overline{X})^2 \quad \text{et} \quad S_2^2 = S_Y^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (Y_i - \overline{Y})^2.$$

- On utilise alors la statistique de test

$$U = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_X^2}{n_1} + \frac{S_Y^2}{n_2}}}.$$

- Sous H_0, $U \approx \mathcal{N}(0, 1)$ approximativement, et on procède de même que précédemment.

c. Cas où les variances σ_1^2 et σ_2^2 sont inconnues mais égales (grands échantillons)

- Si les variances sont inconnues mais égales $\sigma_1^2 = \sigma_2^2 = \sigma^2$, on estime la valeur commune σ^2 par

$$S^2 = \frac{(n_1-1)S_X^2 + (n_2-1)S_Y^2}{n_1+n_2-2} = \frac{1}{n_1+n_2-2}\left(\sum_{i=1}^{n_1}(X_i-\overline{X})^2 + \sum_{i=1}^{n_2}(Y_i-\overline{Y})^2\right).$$

- On montre facilement que l'estimateur obtenu S^2 est sans biais. En effet, S_X^2 est un estimateur sans biais de σ_1^2 donc $\mathbb{E}(S_X^2) = \sigma_1^2 = \sigma^2$, et S_Y^2 est sans biais de σ_2^2 donc $\mathbb{E}(S_Y^2) = \sigma_2^2 = \sigma^2$.
- On en déduit que :

$$\mathbb{E}(S^2) = \frac{(n_1-1)\mathbb{E}(S_X^2) + (n_2-1)\mathbb{E}(S_Y^2)}{n_1+n_2-2} = \frac{(n_1-1)\sigma^2 + (n_2-1)\sigma^2}{n_1+n_2-2} = \sigma^2$$

c. Cas où les variances σ_1^2 et σ_2^2 sont inconnues mais égales (grands échantillons)

- On estime alors $\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)$ par $S^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)$
- On utilise ainsi la statistique de test

$$U = \frac{\overline{X} - \overline{Y}}{\sqrt{S^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}.$$

- Sous H_0, $U \approx \mathcal{N}(0,1)$ approximativement, et on procède de même que précédemment.

Exemple 3

Un fabricant de câbles en acier étudie un nouveau traitement de câbles pour améliorer leur résistance. Il choisit au hasard 200 câbles traités et 100 câbles non traités. On suppose que la charge de rupture est une variable aléatoire. On note X_i la charge de rupture du ième câble traité et Y_i la charge de rupture du ième câble non traité. On observe $\overline{x} = 30,82$: $\overline{y} = 29,63$;

$$\frac{1}{199}\sum_{i=1}^{200}(x_i - \overline{x})^2 = 27,25 \text{ et } \frac{1}{99}\sum_{i=1}^{100}(y_i - \overline{y})^2 = 23,99.$$

Peut-on conclure à l'efficacité du traitement ?

Exemple 3

- Soit μ_1 (respectivement μ_2) la charge de rupture moyenne (dans la population) des câbles traités (respectivement non traités), σ_1^2 (respectivement σ_2) la variance.
- On suppose que les deux échantillons $X_1, \ldots, X_{n_1}$, $n_1 = 200$, et $Y_1, \ldots, Y_{n_2}$, $n_2 = 100$ sont indépendants.
- μ_1 et μ_2 sont estimés par $\overline{X}$ et $\overline{Y}$ les charges moyennes empiriques des câbles traités et non traités des échantillons.
- $\overline{x} = 30.82$ (resp. $\overline{y} = 29.63$) est la réalisation de $\overline{X}$ (resp. $\overline{Y}$).
- Les variances σ_1^2 et σ_2^2 sont inconnues et estimées par les variances empiriques S_X^2 et S_Y^2.
- $s_X^2 = 27.25$ (resp. $s_X^2 = 23.99$) est la réalisation de S_X^2 (resp. S_Y^2) de S_X^2 (resp. S_Y^2).

Exemple 3

- On souhaite tester

$$H_0 \ : \ \mu_1 = \mu_2 \ \text{ contre } H_1 \ : \ \mu_1 > \mu_2.$$

- Le TCL s'applique (les deux échantillons sont suffisamment grands), et la statistique de test est en l'absence d'information sur les variances

$$T = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_X^2}{n_1} + \frac{S_Y^2}{n_2}}} \approx \mathcal{N}(0,1) \quad \textit{sous } H_0$$

Exemple 3

Règle de décision :

- si $T > c_\alpha$ on rejette H_0,
- si $T \leq c_\alpha$, on ne rejette pas H_0,

où le seuil c_α est tel que $\mathbb{P}_{H_0}(T > c_\alpha) = \alpha = 5\%$. Dans la table de la loi $\mathcal{N}(0,1)$, on lit $c_\alpha = 1,645$. Si on note t la réalisation de T sur les deux échantillons, si $t > 1.645$ on rejettera H_0 (avec un risque de 5% de se tromper), si $t \leq 1.645$ on ne rejettera pas H_0.

Mise en oeuvre : Sur l'échantillon, on trouve $t = 1,94 > 1,645$. On rejette H_0 au risque 5%. Le traitement est donc efficace.

3.2.2 Petits échantillons gausssiens

- On considère à nouveau deux échantillons indépendants, le premier $(X_1, \ldots, X_{n_1})$, d'espérance $\mathbb{E}(X_i) = \mu_1$, de variance $\mathrm{Var}(X_i) = \sigma_1^2$, et le second $(Y_1, \ldots, Y_{n_2})$, d'espérance $\mathbb{E}(Y_i) = \mu_2$, de variance $\mathrm{Var}(Y_i) = \sigma_2^2$.
- Si l'un des deux effectifs n_1 ou n_2 n'est pas assez grand pour appliquer le TCL, on ne peut pas utiliser les résultats de la section précédente !
- Cependant, on peut construire un test analogue dans le cas où les deux échantillons sont gaussiens.
- On suppose dans la suite que $X_i \sim \mathcal{N}\left(\mu_1, \sigma_1^2\right)$, $i = 1, \ldots, n_1$ et $Y_i \sim \mathcal{N}\left(\mu_2, \sigma_2^2\right)$, $i = 1, \ldots, n_2$.

Supposons qu'on teste :

$$\mathbf{H_0} \ : \ \mu_1 = \mu_2 \ \text{ contre } \mathbf{H_1} \ : \ \mu_1 \neq \mu_2$$

Le test est encore fondé sur la différence $\overline{X_{n_1}} - \overline{Y_{n_2}}$. Dans le cas gaussien, on connaît la loi exacte de $\overline{X_{n_1}}$ et $\overline{Y_{n_2}}$:

$$\overline{X_{n_1}} \sim \mathcal{N}\left(\mu_1, \frac{\sigma_1^2}{n_1}\right) \quad \text{et} \quad \overline{Y_{n_2}} \sim \mathcal{N}\left(\mu_2, \frac{\sigma_2^2}{n_2}\right).$$

Par indépendance des deux échantillons,

$$\overline{X_{n_1}} - \overline{Y_{n_2}} \sim \mathcal{N}\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right).$$

Sous H_0,

$$\overline{X_{n_1}} - \overline{Y_{n_2}} \sim \mathcal{N}\left(0, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right) \quad \text{soit encore} \quad \frac{\overline{X_{n_1}} - \overline{Y_{n_2}}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \mathcal{N}(0,1).$$

(i) Cas où les variances σ_1^2 et σ_2^2 sont connues

La statistique de test est

$$T_{n_1,n_2} = \frac{\overline{X_{n_1}} - \overline{Y_{n_2}}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}.$$

Sous H_0, $T_{n_1,n_2} \sim \mathcal{N}(0,1)$ et on procède de la même manière que précédemment.

(ii) Cas où les variances σ_1^2 et σ_2^2 sont inconnues mais égales

- On estime la valeur commune σ^2 par la moyenne pondérée par les effectifs $(n_1 - 1)$ et $(n_2 - 1)$ des deux variances d'échantillons S_1^2 et S_2^2 :

$$S^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

de sorte que $\sigma_1^2/n_1 + \sigma_2^2/n_2 = \sigma^2 (1/n_1 + 1/n_2)$ est estimée par $S^2 (1/n_1 + 1/n_2)$.

- On utilise alors la statistique de test

$$T_{n_1,n_2} = \frac{\overline{X_{n_1}} - \overline{Y_{n_2}}}{\sqrt{S^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}.$$

- On montre que dans le cas gaussien, sous H_0,

$$T_{n_1,n_2} \sim St(n_1 + n_2 - 2).$$

(iii) Cas où σ_1^2 et σ_2^2 sont inconnues mais pas forcément égales

- Les solutions ne sont pas satisfaisantes pour des petits échantillons lorsque les variances sont différentes et inconnues.
- Néanmoins, lorsque les deux échantillons sont gaussiens et que les variances sont inconnues et différentes, on peut montrer que la statistique

$$T_{n_1,n_2} = \frac{\overline{X_{n_1}} - \overline{Y_{n_2}}}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

suit, sous H_0, une loi de Student $St(\nu)$ avec ν : l'entier naturel le plus proche du quotient :

$$\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2 \Big/ \left[\frac{(\frac{s_1^2}{n_1})^2}{n_1 - 1} + \frac{(\frac{s_2^2}{n_2})^2}{n_2 - 1}\right].$$

Exemple 4

On cherche à savoir si le rythme cardiaque d'un individu augmente lorsque l'individu est soumis à un stress. Pour cela, on mesure le rythme cardiaque de 5 individus avant une séance de cinéma passant un film d'horreur, et le rythme cardiaque de 5 autres individus après la séance de cinéma. On suppose que le rythme cardiaque d'un individu est une variable aléatoire de loi normale. Les rythmes cardiaques observés sont les suivants :

avant la séance	90	76	80	87	83
après la séance	98	77	88	90	89

1. Le rythme cardiaque augmente t-il de manière significative avec le stress ?
2. Calculer le degré de signification du test.

Exemple 5

On admet que la production de lait d'une vache normande (respectivement hollandaise) est une v.a. de loi $\mathcal{N}(\mu_1, \sigma_1^2)$ (respectivement $\mathcal{N}(\mu_2, \sigma_2^2)$). Un producteur de lait souhaite comparer le rendement des vaches normandes et hollandaises de son unité de production. Les relevés de production de lait (exprimée en litres) de 10 vaches normandes et hollandaises ont donné les résultats suivants :

vaches normandes	552	464	483	506	497	544	486	531	496	501
vaches hollandaises	487	489	470	482	494	500	504	537	482	526

Les deux races de vaches laitières ont-elles le même rendement ?

Application avec R

Si on suppose l'égalité des variances :

$> x < -c(552, 464, 483, 506, 497, 544, 486, 531, 496, 501)$
$> y < -c(487, 489, 470, 482, 494, 500, 504, 537, 482, 526)$
$> t.test(x, y, var.equal = T)$

Two Sample t-test

data : x and y

$$t = 0.8074, \text{df} = 18, \text{p-value} = 0.43$$

alternative hypothesis : true difference in means is not equal to 0
95 percent confidence interval :

$$-14.25774 \quad 32.05774$$

sample estimates : mean of x mean of y
506.0 497.1

Application avec R

Si on ne suppose pas l'égalité des variances :

$> x <- c(552, 464, 483, 506, 497, 544, 486, 531, 496, 501)$
$> y <- c(487, 489, 470, 482, 494, 500, 504, 537, 482, 526)$
$> t.test(x, y, var.equal = F)$

Welch Two Sample t-test

data : x and y

$t = 0.8074$, df $= 16.553$, p-value $= 0.4309$

alternative hypothesis : true difference in means is not equal to 0
95 percent confidence interval :

$-14.40370 \quad 32.20370$

sample estimates : mean of x mean of y
506.0 497.1
Conclusion : Au risque 5%, le test n'est pas significatif.

Utilisation en pratique

- En pratique on préférera utiliser la version asymptotique du test si la taille d'échantillon le permet.
- Si l'échantillon est trop petit on utilisera généralement un test **nonparamétrique** pour compararer les distributions.
- En pratique on ne sait jamais si les variances des deux groupes sont égales et donc on utilisera toujours le test de Welch à **variances inégales** !
- Il ne faut pas "enchainer" les tests : tester d'abord l'égalité des variances pour ensuite décider si l'on fait le test de comparaison des espérances à variances égales ou non est une **mauvaise pratique** car la décision finale n'aura pas le bon risque de première espèce ! !

3.3 Test de comparaison de deux proportions

- On souhaite comparer deux proportions p_1 et p_2 d'individus possèdant un même caractère dans deux populations différentes.
- Par exemple, p_1 représente la proportion de favorables à un candidat dans une ville V1, et p_2 la proportion de favorables à ce candidat dans une autre ville V2
- On peut se demander si la proportion de favorables au candidat est la même dans les deux villes, auquel cas on testera

$$H_0 : p_1 = p_2 \text{ contre } H_1 : p_1 \neq p_2$$

- On va donc considérer deux échantillons indépendants $(X_1, \cdots, X_{n_1})$ de loi $\mathcal{B}(p_1)$, et $(Y_1, \cdots, Y_{n_2})$ de loi $\mathcal{B}(p_2)$.
- Dans le cas de l'exemple, X_i représente l'opinion du ième électeur dans V1, Y_i dans V2.
- La proportion théorique p_1 (resp. p_2) est estimée par la proportion aléatoire $n_1^{-1} \sum_{i=1}^{n_1} X_i = \overline{X_{n_1}}$ (resp. $n_2^{-1} \sum_{i=1}^{n_2} Y_i = \overline{Y_{n_2}}$).

- On rappelle que $\mathbb{E}(X_i) = p_1$, $\mathrm{Var}(X_i) = p_1(1-p_1)$, $\mathbb{E}(Y_i) = p_2$, $\mathrm{Var}(Y_i) = p_2(1-p_2)$, $\mathbb{E}(\overline{X_{n_1}}) = p_1$, $\mathrm{Var}(\overline{X_{n_1}}) = p_1(1-p_1)/n_1$, $\mathbb{E}(\overline{Y_{n_2}}) = p_2$ et $\mathrm{Var}(\overline{Y_{n_2}}) = p_2(1-p_2)/n_2$.
- Le test est fondé sur l'écart $\overline{X_{n_1}} - \overline{Y_{n_2}}$ entre les deux proportions aléatoires observées dans les échantillons, v.a. dont il faut connaître la loi sous H_0.
- Si les tailles d'échantillons n_1 et n_2 sont suffisamment importantes, le **TCL** s'applique et on a

$$\overline{X_{n_1}} \approx \mathcal{N}\left(p_1, \frac{p_1(1-p_1)}{n_1}\right) \quad \text{et} \quad \overline{Y_{n_2}} \approx \mathcal{N}\left(p_2, \frac{p_2(1-p_2)}{n_2}\right)$$

- soit encore, sous H_0

$$\frac{\overline{X_{n_1}} - \overline{Y_{n_2}}}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \approx \mathcal{N}(0,1)$$

- On connaît la loi de cette v.a. sous H_0, mais on ne peut l'utiliser comme statistique de test car on ne connaît pas la valeur de $p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2$. Il faut donc estimer cette quantité sous H_0.
- Sous H_0, $p_1 = p_2 = p$, et on estime la valeur commune p par

$$P_n = \frac{\sum_{i=1}^{n_1} X_i + \sum_{i=1}^{n_2} Y_i}{n_1 + n_2} = \frac{n_1 \overline{X_{n_1}} + n_2 \overline{Y_{n_2}}}{n_1 + n_2}.$$

  P_n est la proportion aléatoire observée sur les deux échantillons (on note n la taille d'échantillon globale, $n = n_1 + n_2$).
- On estime alors (sous H_0) $p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2 = p(1-p)(1/n_1 + 1/n_2)$ par $P_n(1-P_n)(1/n_1 + 1/n_2)$. On construit le test à partir de la statistique

$$T_{n_1,n_2} = \frac{\overline{X_{n_1}} - \overline{Y_{n_2}}}{\sqrt{P_n(1-P_n)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Sous les hypothèses suivantes :

- $n = n_1 + n_2$ suffisamment "grand", en pratique $n \geq 30$,
- $n_1 p_n \geq 5, n_1(1-p_n) \geq 5$ et $n_2 p_n \geq 5, n_2(1-p_n) \geq 5$, où p_n est la réalisation de P_n,

alors $T_{n_1,n_2} \approx \mathcal{N}(0,1)$.

Règle de décision :

- si $|T_{n_1,n_2}| > c_\alpha$ on rejette H_0,
- si $|T_{n_1,n_2}| \leq c_\alpha$, on ne rejette pas H_0,

où le seuil c_α est choisi tel que $\mathbb{P}_{H_0}(|T_{n_1,n_2}| > c_\alpha) = \alpha$; la valeur de c_α est lue dans la table de la $\mathcal{N}(0,1)$.

Exemple 6

A la sortie de deux salles de cinéma donnant le même film, on a interrogé des spectateurs quant à leur opinion sur le film. Les résultats de ce sondage d'opinion sont les suivants :

Opinion	Mauvais film	Bon film	Total
Salle 1	30	70	100
Salle 2	48	52	100
Total	78	122	200

L'opinion est-elle significativement liée à la salle ?

Soit p_1 (resp. p_2) la proportion de gens de la salle 1 (resp. salle 2) ayant une mauvaise opinion du film.

- On veut tester $H_0 : p_1 = p_2$ contre $H_1 : p_1 \neq p_2$.
- Soit X_i (resp. Y_i) la v.a. représentant l'opinion du ième spectateur interrogé dans la salle 1 (resp. salle 2). $X_i = 1$ si le ième spectateur interrogé dans la salle 1 a une mauvaise opinion, $X_i = 0$ sinon.
- $X_i \sim \mathcal{B}(p_1)$ et $Y_i \sim \mathcal{B}(p_2)$.
- On note $\overline{X_{n_1}}$ (resp. $\overline{Y_{n_2}}$) la proportion empirique de spectateurs interrogés dans la salle 1 (resp. salle 2) ayant une mauvaise opinion.
- Les effectifs $n_1 = n_2 = 100$ sont suffisamment grands pour appliquer le TCL.
- La statistique de test est

$$T_{n_1,n_2} = \frac{\overline{X_n} - \overline{Y_n}}{\sqrt{P_n(1-P_n)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \approx \mathcal{N}(0,1) \text{ sous } H_0$$

où $P_n = (n_1\overline{X_{n_1}} + n_2\overline{Y_{n_2}})/(n_1 + n_2)$.

Règle de décision :

- si $|T_{n_1,n_2}| > c_\alpha$ on rejette H_0,
- si $|T_{n_1,n_2}| \leq c_\alpha$, on ne rejette pas H_0,

où le seuil c_α est choisi tel que $\mathbb{P}_{H_0}(|T_{n_1,n_2}| > c_\alpha) = \alpha = 5\%$; la valeur de $c_\alpha = 1,96$ est lue dans la table de la loi $\mathcal{N}(0,1)$.

Mise en oeuvre : soient $\overline{x_{n_1}}$, $\overline{y_{n_2}}$, p_n les réalisations de $\overline{X_{n_1}}$, $\overline{Y_{n_2}}$, P_n.
$\overline{x_{n_1}} = 30/100$, $\overline{y_{n_2}} = 48/100$, $p_n = (30+48)/200$.
Ainsi, $t_n = (\overline{x_{n_1}} - \overline{y_{n_2}})/\sqrt{p_n(1-p_n)(1/100+1/100)} = -2.61$.
$|t_n| = 2.61 > 1.96 = c_\alpha$: il y a une différence significative entre les deux salles au niveau 5%.
Degré de signification :

$$\text{pvaleur} = \mathbb{P}_{H_0}(|T_{n_1,n_2}| > 2.61) = 2(1 - \mathbb{P}_{H_0}(T \leq 2.61)) = 2(1-0.995) = 0.01.$$

3.4 Test de comparaison de deux variances (cas gaussien) : test de Fisher

- On souhaite comparer les variances σ_1^2 et σ_2^2 d'un même caractère dans deux populations différentes.
- **Le test de Fisher ne fonctionne que si le caractère est distribué dans les deux populations suivant une loi normale.**
- On considère deux échantillons gaussiens $(X_1, \ldots, X_{n_1})$, $X_i \sim \mathcal{N}(\mu_1, \sigma_1)$ et $(Y_1, \ldots, Y_{n_2})$, $Y_i \sim \mathcal{N}(\mu_2, \sigma_2)$.
- Soient $\overline{X_{n_1}} = \sum_{i=1}^{n_1} X_i/n_1$ estimateur de μ_1, $\overline{Y_{n_2}} = \sum_{i=1}^{n_2} Y_i/n_2$ estimateur de μ_2,

$$S_1^2 = \sum_{i=1}^{n_1} (X_i - \overline{X_{n_1}})^2/(n_1 - 1)$$

estimateur de σ_1^2, et

$$S_2^2 = \sum_{i=1}^{n_2} (Y_i - \overline{Y_{n_2}})^2/(n_2 - 1)$$

estimateur de σ_2^2.

- On veut comparer σ_1^2 et σ_2^2 : à partir des variances observées S_1^2 et S_2^2, peut-on dire si $\sigma_1^2 = \sigma_2^2$?

On va donc construire le test de $H_0 : \sigma_1^2 = \sigma_2^2$ contre $H_1 : \sigma_1^2 \neq \sigma_2^2$.

- Sous H_0,
$$\sigma_1^2 = \sigma_2^2, \text{ i.e. } \sigma_1^2/\sigma_2^2 = 1.$$
- Le test va alors être fondé sur le rapport S_1^2/S_2^2 que l'on comparera à la valeur de référence 1.
- Si ce rapport est significativement différent de 1, on rejettera H_0, sinon on ne rejettera pas H_0.
- Quelle est la loi de la statistique de test $T = S_1^2/S_2^2$ sous H_0 ?

Loi de la statistique de test $T = S_1^2/S_2^2$ sous H_0

Rappel : Loi de Fisher

- Soient X, Y deux v.a. indépendantes telles que $X \sim \chi^2(d_1)$ et $Y \sim \chi^2(d_2)$. Alors la v.a.

$$\mathbf{(X/d_1)/(Y/d_2) \sim \mathcal{F}(d_1, d_2)},$$

  $\mathcal{F}(d_1, d_2)$ étant la loi de de Fisher à d_1 et d_2 degrés de liberté.
- La loi de Fisher est une loi non symétrique et une v.a. de Fisher ne prend que des valeurs positives.

Loi de la statistique de test $T = S_1^2/S_2^2$ sous H_0

- Puisque les échantillons sont gaussiens, on sait que

$$\frac{(n_1-1)S_1^2}{\sigma_1^2} \sim \chi^2(n_1-1) \quad \text{et} \quad \frac{(n_2-1)S_2^2}{\sigma_2^2} \sim \chi^2(n_2-1).$$

- De plus, ces deux v.a. sont indépendantes.
- On a donc

$$\frac{\frac{(n_1-1)S_1^2}{\sigma_1^2}/(n_1-1)}{\frac{(n_2-1)S_2^2}{\sigma_2^2}/(n_2-1)} \sim \frac{\chi^2(n_1-1)/(n_1-1)}{\chi^2(n_2-1)/(n_2-1)} \sim \mathcal{F}(n_1-1, n_2-1).$$

- Or sous H_0, $\sigma_1^2 = \sigma_2^2$, donc $T = S_1^2/S_2^2 \sim \mathcal{F}(n_1-1, n_2-1)$ sous H_0.

Règle de décision

- si $T < c_\alpha$ $(c_\alpha < 1)$ ou si $T > d_\alpha$ $(d_\alpha > 1)$ on rejette H_0,
- si $c_\alpha \leq T \leq d_\alpha$, on ne rejette pas H_0.

Les seuils c_α et d_α sont choisis tels que :

$$\mathbb{P}_{H_0}(\text{rejeter } H_0) = \mathbb{P}_{H_0}(T < c_\alpha \text{ ou } T > d_\alpha) = \mathbb{P}_{H_0}(T < c_\alpha) + \mathbb{P}_{H_0}(T > d_\alpha) = \alpha$$

avec α fixé à l'avance.

- Or **sous** H_0, $T \sim \mathcal{F}(n_1 - 1, n_2 - 1)$, donc les valeurs de c_α et d_α sont lues dans la table de la Fisher $\mathcal{F}(n_1 - 1, n_2 - 1)$ tels que

$$\mathbb{P}(\mathcal{F}(n_1 - 1, n_2 - 1) < c_\alpha) = \alpha/2$$

et

$$\mathbb{P}(\mathcal{F}(n_1 - 1, n_2 - 1) > d_\alpha) = \alpha/2,$$

c.à.d

$$\mathbb{P}(\mathcal{F}(n_1 - 1, n_2 - 1) \leq d_\alpha) = 1 - \alpha/2.$$

Remarques

- Il y a donc deux bornes à lire dans la table de la Fisher.
- La valeur de d_α (> 1) se lit directement dans la table de $\mathcal{F}(n_1 - 1, n_2 - 1)$.
- En revanche, la lecture de c_α pose plus de problème car $c_\alpha < 1$.
- Mais on rappelle que si $F \sim \mathcal{F}(d_1, d_2)$, alors $\frac{1}{F} \sim \mathcal{F}(d_2, d_1)$, et donc pour lire une valeur plus petite que 1, il suffit de prendre l'inverse de la valeur lue dans la table $\mathcal{F}(d_2, d_1)$ où l'on a inversé les degrés de liberté.
- Ainsi, si $F \sim \mathcal{F}(n_1 - 1, n_2 - 1)$ et si $c_\alpha < 1$,

$$\mathbb{P}(F < c_\alpha) = \mathbb{P}(1/F > 1/c_\alpha) = \alpha/2$$

et

$$\mathbb{P}(1/F \leq 1/c_\alpha) = 1 - \alpha/2.$$

$1/c_\alpha$ est égal au quantile d'ordre $1 - \alpha/2$ de la loi $\mathcal{F}(n_2 - 1, n_1 - 1)$.

Exemples

1. Tester l'égalité des variances dans l'exemple du test de comparaison de moyennes du cas gaussien sur les vaches hollandaises et normandes.
2. Dans deux Unités d'Enseignement et de Recherche (UER), $U1$ et $U2$, de psychologie, on suppose que les notes de statistiques des étudiants suivent des lois normales. On observe les résultats suivants sur un échantillon de 25 étudiants de l'UER $U1$ et de 10 étudiants de l'UER $U2$:

$$\sum_{i=1}^{25} x_i = 310 \quad \sum_{i=1}^{25} x_i^2 = 3916 \quad \sum_{i=1}^{10} y_i = 129 \quad \text{et} \quad \sum_{i=1}^{10} y_i^2 = 1709.1$$

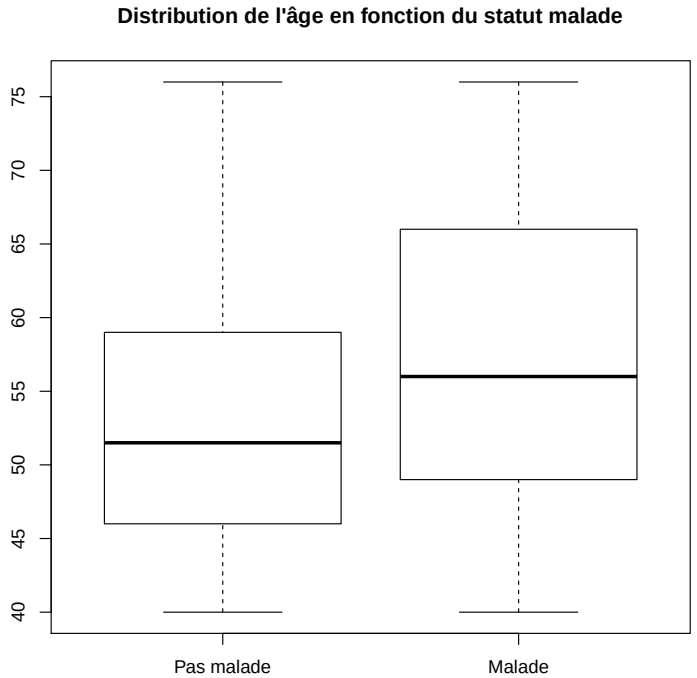
où x_i (respectivement y_i) désigne la note obtenue par le $i^{ème}$ étudiant de l'échantillon de $U1$ (respectivement $U2$).
Peut-on dire que les variances des notes de statistiques dans les deux UER sont significativement différentes ? (on prendra un risque de 5%).

3.5 Test de comparaison de deux variances dans le cas non-gaussien

- Le test de Fisher n'est valide que si les variables aléatoires X et Y sont gaussiennes. Il n'y a pas de résultat asymptotique pour le test de Fisher !
- En pratique ce test est rarement utilisé à cause des trop fortes hypothèses nécessaires sur la loi des échantillons.
- Il existe d'autres tests plus robustes à la non normalité des échantillons : les tests de Bartlett et de Levene.
- Le test de Bartlett est un peu moins sensible à la non-normalité des échantillons que le test de Fisher.
- Le test de Levene est asymptotiquement valide même si les échantillons sont non gaussiens.
- En pratique, quand les tailles d'échantillons n_1 et n_2 sont grands on utilisera toujours le test de Levene sans se soucier de la normalité des échantillons.

Données Evans

- La base de données Evans contient les informations sur une cohorte de 609 hommes ayant été suivis sur une période de 7 ans. Le but de l'étude est d'étudier les facteurs de risque de l'apparition d'une maladie cardiaque des coronaires.
- Nous nous intéresserons uniquement à comparer la distribution de l'âge (au début de l'étude) en fonction de la présence ou pas d'une maladie cardiaque des coronaires chez les patients. L'âge est-il le même chez les patients malades que chez les non malades ? En terme d'espérance ? En terme de variance ?
- Soit $X_1, \ldots, X_{n_1}$ (resp. $Y_1, \ldots, Y_{n_2}$) l'âge des patients non malades (resp. malades) d'espérance inconnue μ_1 (resp. μ_2) et de variance inconnue σ_1^2 (resp. σ_2^2). On suppose X_i indépendant de Y_j, $i = 1, \ldots, n_1$, $j = 1, \ldots, n_2$.
- $n_1 = 538$ et $n_2 = 71$; on est dans le cas de grand échantillons.



- $\overline{x_{n_1}} = 53.23$, $\overline{y_{n_2}} = 57.25$, $s_1 = 9.04$, $s_2 = 10.13$.
- D'un point de vue descriptif les distributions semblent différentes. Les patients malades ont tendance à être plus âgés que les patients non malades. La dispersion semble être légèrement plus grande chez les malades également.

Test de comparaison d'espérance sous R

- > t.test(age∼chd)
  Welch Two Sample t-test
  data : age by chd t = -3.1784, df = 85.392, p-value = 0.002063 alternative hypothesis : true difference in means is not equal to 0
- Le test est extrêmement significatif : les espérances d'âge diffèrent entre les deux groupes !

Test de comparaison de variance sous R

- > levene.test(age,chd)
  modified robust Brown-Forsythe Levene-type test based on the absolute deviations from the median
  data : age Test Statistic = 3.7765, p-value = 0.05244
- Même si le test n'est pas significatif à 5% il semblerait que les variances ne soient pas les mêmes dans les deux groupes. On peut dire que les variances semblent inégales avec 5.24% de chances environ de se tromper en disant cela.

Remarques

- On ne peut pas faire le test de Fisher ici car les distributions des deux échantillons ne semblent pas gaussiennes.
- Le test de Fisher ou de Bartlett renvoie des résultats faux ici !
- $>$ var.test(age$\sim$chd)

  F test to compare two variances

  data : age by chd F $=$ 0.7978, num df $=$ 537, denom df $=$ 70, p-value $=$ 0.1808
- $>$ bartlett.test(age$\sim$chd)

  Bartlett test of homogeneity of variances

  data : age by chd Bartlett's K-squared $=$ 1.6662, df $=$ 1, p-value $=$ 0.1968