Test à 2 échantillons

Table of contents

1 Introduction

1.1 Dans les cours précédents

— On comparait l'éspérance d'un échantillon à une valeurs théorique.

— *On faisait :*
  — Des Hypothèses
  — Une statisitque de test et sa loi sous H_0
  — Un zone de rejet
  — Une p-valeur En fonction de la p-valeur (ou de la zone de rejet) on pouvait conclure sur le rejet (ou non) de H_0.
— *On maitrisait :* Le risque de première espèce (la probabilité de rejeté H_0 à tort)
— On pouvait également calculer la puissance (la probabilité de rejeter H_0 à raison pour un écart à l'espérance sous H_0 donné).

Maintenant : C'est tout pareil !

1.2 Dans ce cours

On souhaite comparer deux échantillons (potentiellement issue de 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

La généralisation à plus de deux populations sera faite dans le cadre du cours de l'analyse de la variance (modèle linéaire).

2 Test de comparaison de deux espérances sur échantillons appariés

2.1 Description du problème (échantillons appariés)

— Modélisation : on a d'une part $X_1, ..., X_n$ i.i.d distribuées selon une loi d'espérance μ_1 et d'autre part, $Y_1, \dots, Y_n$ 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).
— On s'intéressera donc 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 vues en cours de test à un échantillon sur les Z_j.

2.2 Exemple 1 : Rythme cardiaque avant et après la prise d'un médicament

Exemple 1

Dans 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 significatif 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 ?

Pour résoudre ce problème on se ramène à ce qu'on a vu précedemment :

Sortie R :

```
x <- c(80,80,82,75,80,74,80,72,91,88,70,65,83,74,81,68,69,
       71,70,73,78,75,76,78,77,75,72,71,75,78)
y <- c(85,84,87,81,79,85,87,78,96,80,82,73,89,85,86,72,
       74,77,75,81,70,77,76,82,83,80,80,81,76,77)
d <- x - y
mean(d)
```

```
[1] -4.566667
```

```
sd(d)
```

```
[1] 4.695437
```

```
t.test(d, mu = 0, alternatives ="two.sided")
```

```
    One Sample t-test

data:  d
t = -5.327, df = 29, p-value = 1.022e-05
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -6.319972 -2.813362
sample estimates:
mean of x
-4.566667
```

```r
## ou encore :
t.test(x, y, paired= TRUE, mu = 0, alternatives ="two.sided")
```

```
    Paired t-test

data:  x and y
t = -5.327, df = 29, p-value = 1.022e-05
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 -6.319972 -2.813362
sample estimates:
mean difference
      -4.566667
```

3 Test de comparaison de deux espérances sur échantillons indépendants

3.1 Description des données

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 consiste à évaluer l'efficacité d'un médicament en le comparant avec un placebo :

- Des exercices respiratoires sont prescrits à 67 patients choisis au hasard. Ces patients prennent tous un placebo (goupe témoin (A)).
- Les mêmes exercices respiratoires sont prescrits aux 70 autres patients. Ces patients prennent tous 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 par un médecin (voir tableau suivant) sur une échelle de 0 (pas d'amélioration) à 10 (récupération totale).

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	$n_1 = 67$	$n_2 = 70$
	$\bar{x} = 4,776$	$\bar{y} = 5,671$
	$s_1^2 = 7,479$	$s_2^2 = 2,601$

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

Remarque sur l'exemple :

- L'amélioration moyenne du groupe traité est supérieure de 0.9 environ 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 mesure l'amélioration de la capacité pulmonaire associe pour un patient du groupe A. L'espérance μ_1 et l'écart-type σ_1 de X sont inconnus.
  - Soit Y qui mesure l'amélioration de la capacité pulmonaire associe pour un patient du groupe B. 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.

Plus généralement :

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. $(X_1, \dots, X_{n_1})$ de taille n_1, extrait de cette population et tels que $\mathbb{E}(X_i) = \mu_1$ et $\mathrm{V}ar(X_i) = \sigma_1^2$; on note $\overline{X}$ la moyenne empirique et S_1^2 la variance empirique de cet échantillon càd :

$$\overline{X} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i \text{ et } S_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(X_i - \overline{X})^2.$$

— 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 $\mathrm{V}ar(Y_i) = \sigma_2^2$; on note $\overline{Y}$ la moyenne empirique et S_2^2 la variance empirique de cet échantillon càd :

$$\overline{Y} = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i \text{ et } S_2^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(Y_i - \overline{Y})^2.$$

— les variables X_i et Y_j sont supposées indépendantes (mesures réalisées sur des individus nécessairement différents).

3.2 Les hypothèses

A l'aide des deux échantillons on veut comparer ces deux populations. 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 $H_1 \ : \ \mu_1 > \mu_2$.

Dans l'Exemple 2

— En pratique on cherche toujours à contrôler la probabilité de décider à tort que le médicament est plus efficace que le placebo (point de vue de l'agence sanitaire).
— 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.

3.3 Statistiques de test

Comme précédement la loi de la statistique de test va dépendre :

- De la taille des échantillons
- De si on connait la variance ou non
- D'une hypothèse de normalité ou non.

De plus : elle va dépendre de si la variance est supposée égale dans les deux échantillons !

Cadre de grands échantillons (loi quelconque)

Dans cette partie nous allons utiliser le TCL : 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 $\overline{Y}$. La théorie nous dit qu'il faut avoir $n_1 \geq 30$ et $n_2 \geq 30$ pour pouvoir utiliser les résultats asymptotiques.

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.

Premier cas : grand échantillon variances connues :

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

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

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

$$\overline{X} - \overline{Y} \underset{H_0}{\approx} N\left(0, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right).$$

- Soit encore, :

$$T_{n_1,n_2} = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \underset{H_0}{\approx} N(0,1).$$

- Mais on ne peut utiliser directement la v.a. T_{n_1,n_2} comme statistique de test que lorsque les variances des deux populations sont connues.

Deuxième cas : grand échantillon variances inconnues (pas égales) :

— 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

$$T_{n_1,n_2} = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_X^2}{n_1} + \frac{S_Y^2}{n_2}}} \underset{H_0}{\approx} N(0,1)$$

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

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

(Cette variance est la moyenne pondérée, par les tailles d'échantillons des deux variances précédentes)

— 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 :

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

— On estime alors

$$\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \text{ par } S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

— On utilise ainsi la statistique de test

$$T_{n_1,n_2} = \frac{\overline{X} - \overline{Y}}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \underset{H_0}{\approx} N(0,1)$$

Cadre petit échantillons gaussiens

— On considère à nouveau deux échantillons indépendants, le premier $(X_1, \dots, X_{n_1})$, d'espérance $\mathbb{E}(X_i) = \mu_1$, de variance $\mathrm{V}ar(X_i) = \sigma_1^2$, et le second $(Y_1, \dots, Y_{n_2})$, d'espérance $\mathbb{E}(Y_i) = \mu_2$, de variance $\mathrm{V}ar(Y_i) = \sigma_2^2$.
— 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 $X_i \sim N(\mu_1, \sigma_1^2)$, $i = 1, \dots, n_1$ et $Y_i \sim N(\mu_2, \sigma_2^2)$, $i = 1, \dots, n_2$.

A nouveau on a trois cas : 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.

Cadre petit échantillons gaussiens : variances connues

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 $\overline{Y_{n_2}}$:

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

Par indépendance des deux échantillons,

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

Sous H_0,

$$\overline{X_{n_1}} - \overline{Y_{n_2}} \underset{H_0}{\sim} N\left(0, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right) \quad \text{soit encore} \quad \frac{\overline{X_{n_1}} - \overline{Y_{n_2}}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \underset{H_0}{\sim} N(0,1).$$

La statistique de test est

$$T_{n_1,n_2} = \frac{\overline{X_{n_1}} - \overline{Y_{n_2}}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \underset{H_0}{\sim} N(0,1)$$

Cas où les variances 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

$$T_{n_1,n_2} = \frac{\overline{X_{n_1}} - \overline{Y_{n_2}}}{\sqrt{S^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \underset{H_0}{\sim} St(n_1+n_2-2).$$

Cas où les variances sont inconnues mais pas forcément égales

Lorsque les deux échantillons sont gaussiens et que les variances sont inconnues et différentes, on peut montrer que la statistique

$$T_{n_1,n_2} = \frac{\overline{X_{n_1}} - \overline{Y_{n_2}}}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}} \underset{H_0}{\sim} St(\nu)$$

avec ν, l'entier naturel le plus proche du quotient :

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

(Admis et la valeur du degrés de liberté n'est pas à connaitre)

Régle de décision, conclusion

Exemple 3

On admet que la production de lait d'une vache normande (respectivement hollandaise) est une v.a. de loi $N(\mu_1, \sigma_1^2)$ (respectivement $N(\mu_2, \sigma_2^2)$). 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 ?

```
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 = 0.80743, df = 18, p-value = 0.43
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -14.25774  32.05774
sample estimates:
mean of x mean of y
    506.0     497.1
```

```
t.test(x,y,var.equal=F)
```

```
    Welch Two Sample t-test

data:  x and y
t = 0.80743, df = 16.553, p-value = 0.4309
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -14.4037  32.2037
sample estimates:
mean of x mean of y
    506.0     497.1
```

Remarques importantes

- En pratique on préférera utiliser la version asymptotique du test (section 3.2.1) si les tailles d'échantillon le permettent.
- Si l'un des deux échantillons est trop petit on utilisera généralement un test **n**onparamétrique pour comparer les distributions (sera vu en BUT3).
- Alternativement on peut utiliser le test dans le cas gaussien (section 3.2.2). Pour cela, il faut que d'autres études aient montré que les deux variables (X et Y) sont bien gaussiennes. Ce qui est rare et souvent sujet à débat.
- La règle $n_1 \geq 30$ et $n_2 \geq 30$ pour appliquer le test asymptotique est arbitraire ! En pratique, on peut vouloir imposer une règle plus stricte : par exemple pour un test univarié j'utilise personnellement la règle $n \geq 50$ et pour un test bivarié j'utilise $n_1 \geq 50$ et $n_2 \geq 50$. Si ces conditions ne sont pas vérifiées j'opte pour un test non paramétrique.
- En pratique on ne sait jamais si les variances des deux groupes sont égales et donc on utilisera toujours le test à **v**ariances inégales ! On fait des tests à variances égales dans les exercices de TD, sur les petits échantillons, pour pouvoir faire les calculs car les degrés de liberté du test de Welch sont pénibles à calculer. Mais en pratique on suppose toujours les variances inégales, les logiciels n'ont pas de problèmes pour faire les calculs.
- 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 **m**auvaise pratique (problème de test multiples) car la décision finale n'aura pas le bon risque de première espèce !!
- De la même manière, tester si les échantillons sont gaussiens (avec un test de Shapiro-Wilks par exemple) pour pouvoir appliquer le test de comparaison d'espérance dans le cas gaussien est aussi une **m**auvaise pratique (problème de test multiples).

4 Test de proportion

4.1 Objectifs

On souhaite comparer deux proportions p_1 et p_2 d'individus possédant un même caractère dans deux populations différentes.

> Exemple 3
>
> Soit p_1 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 si la proportion de favorables au candidat est significativement différentes dans les deux villes.

Plus généralement on va donc considérer deux échantillons indépendants $(X_1, \cdots, X_{n_1})$ de loi $B(p_1)$, et $(Y_1, \cdots, Y_{n_2})$ de loi $B(p_2)$.

Comme d'habitude en test on fait

- Hypothèse
- Statistique de test
- Zone de rejet (p-valeur)
- Conclustion

Hypothèses

Selon les cas on va tester :

$H_0 : p_1 = p_2$ vs $H_1 : p_1 \neq p_2$ ou $H_1 : p_1 > p_2$ ou encore $H_1 : p_1 < p_2$

Statistique de test

On veut créer un stat de test (qui a une certaine loi sous H_0)

- Dans le cas de l'exemple, X_i représente l'opinion du ième électeur dans V1 (1 si l'individu i de la V1 vote pour le candidat et 0 sinon), Y_i dans V2.
- La proportion théorique p_1 est estimée par la proportion aléatoire

— La proportion théorique p_2 est estimée par la proportion aléatoire

— On rappelle que $\mathbb{E}(X_i) = p_1$, $\mathrm{V}ar(X_i) = p_1(1-p_1)$, $\mathbb{E}(Y_i) = p_2$, $\mathrm{V}ar(Y_i) = p_2(1-p_2)$, $\mathbb{E}(\overline{X_{n_1}}) = p_1$, $\mathrm{V}ar(\overline{X_{n_1}}) = p_1(1-p_1)/n_1$, $\mathbb{E}(\overline{Y_{n_2}}) = p_2$ et $\mathrm{V}ar(\overline{Y_{n_2}}) = p_2(1-p_2)/n_2$.
— 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 **T**CL s'applique et on a

$$\overline{X_{n_1}} \approx N\left(\quad , \quad \right) \quad \text{et} \quad \overline{Y_{n_2}} \approx N\left(\quad , \quad \right)$$

— soit encore, sous H_0

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

— 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 théorie $n \geq 30$),
— $n_1 p_n \geq 5, n_1(1-p_n) \geq 5$ et $n_2 p_n \geq 5, n_2(1-p_n) \geq 5$, où p_n est la réalisation de P_n,

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

Zone de rejet et p-valeur

Selon les cas :

H_1	$p_1 \neq p_2$	$p_1 > p_2$	$p_1 < p_2$
Rejet	$\lvert T_{n_1,n_2}\rvert > q_{1-\alpha/2}$	$T_{n_1,n_2} > q_{1-\alpha}$	$T_{n_1,n_2} < q_{\alpha}$
p-valeur	$2\mathbb{P}_{H_0}(T_{n_1,n_2} > \lvert t_n\rvert)$	$\mathbb{P}_{H_0}(T_{n_1,n_2} > t_n)$	$\mathbb{P}_{H_0}(T_{n_1,n_2} < t_n)$

où q_α est le quantile α d'une $\mathcal{N}(0,1)$

Exemple 4

A la sortie de deux salles de cinéma donnant le m^eme 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 ?

Remarques :

- Les conditions $n_1 p_n \geq 5, n_1(1-p_n) \geq 5$ et $n_2 p_n \geq 5, n_2(1-p_n) \geq 5$ sont primordiales et doivent toujours être vérifiées.
- La condition $n = n_1 + n_2 \geq 30$ est arbitraire et optimiste.
- Si les conditions d'application du test de proportion ne sont pas vérifiées, on pourra opter pour le test de Fisher exact. Ce test est exact et n'a aucune condition d'application. Il peut cependant être long à implémenter en pratique si la taille d'échantillon n est grande.
- Sous R, le test de proportion s'implémente en utilisant la commande `prop.test`
- Sous R, le test de Fisher exact s'implémente en utilisant la commande `fisher.test`

```
tabcont=matrix(c(30,48,70,52),ncol=2)
tabcont
```

```
     [,1] [,2]
[1,]   30   70
[2,]   48   52
```

```r
prop.test(tabcont)
```

```
    2-sample test for equality of proportions with continuity correction

data:  tabcont
X-squared = 6.074, df = 1, p-value = 0.01372
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.32287342 -0.03712658
sample estimates:
prop 1 prop 2
  0.30   0.48
```

```r
fisher.test(tabcont)
```

```
    Fisher's Exact Test for Count Data

data:  tabcont
p-value = 0.01348
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.2488590 0.8626843
sample estimates:
odds ratio
 0.4661054
```