Démarche Statistique 1

Tests usuels sur échantillons indépendants

Pierre Neuvial, http://stat.genopole.cnrs.fr/members/pneuvial/demstat
Evry, M1 SGO, automne 2014

Echantillons indépendants

Définition

Deux séries d'observations:

- un échantillon de taille n_1: $(X_1, X_2 \ldots X_{n_1})$
- un échantillon de taille n_2: $(Y_1, Y_2 \ldots Y_{n_2})$

Question: la moyenne des deux populations est-elle identique ?

Exemples

- Efficacité d'un nouveau traitement: pour comparer avec l'ancien traitement, la moitié des patients d'un essai clinique sont traités avec l'ancien traitement, l'autre moitié avec le nouveau. Le choix de quel patient est traité avec quel traitement est effectué par tirage au sort.
- Un gène d'intérêt a-t-il un niveau d'expression différent dans deux types de cancers du sein ?

Plan

Tests de comparaison de moyennes

1. Variance connue: **test Z** (rarement appliqué car la variance est généralement inconnue)
2. Variance inconnue: **test de Student**
   - outil: test d'égalité des variances
   - variance identique dans les deux classes
   - variance différente dans les deux classes

Test de Wilcoxon

Test de comparaison de moyennes

Hypothèses et notations

Echantillons

- $(X_1, X_2 \dots X_{n_1})$ indépendantes, identiquement distribuées de loi $\mathcal{N}(\mu_1, \sigma_1)$
- $(Y_1, Y_2 \dots Y_{n_2})$ indépendantes, identiquement distribuées de loi $\mathcal{N}(\mu_2, \sigma_2)$
- les deux échantillons sont indépendants l'un de l'autre

Moyennes empiriques

$$\bar{X} = \frac{1}{n_1} \sum_{i=1}^{n_1} X_i$$

$$\bar{Y} = \frac{1}{n_2} \sum_{j=1}^{n_2} Y_j$$

Variances empiriques

$$S_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_i - \bar{X})^2$$

$$S_2^2 = \frac{1}{n_2 - 1} \sum_{j=1}^{n_2} (Y_j - \bar{Y})^2$$

Propriétés

Moyennes empiriques

- $\bar{X} \sim \mathcal{N}(\mu_1, \frac{\sigma_1^2}{n_1})$
- $\bar{Y} \sim \mathcal{N}(\mu_2, \frac{\sigma_2^2}{n_2})$

$\bar{X}$ et $\bar{Y}$ sont indépendants, donc $\bar{X} - \bar{Y} \sim \mathcal{N}(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$

Loi sous l'hypothèse nulle $H_0 : \mu_1 = \mu_2$

Sous H_0, la loi de $\bar{X} - \bar{Y}$ est donc $\mathcal{N}(0, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$, qui ne dépend que des variances des deux échantillons.

- Si σ_1 et σ_2 sont connus: **test Z**
- Si σ_1 et σ_2 sont inconnus: **test de Student**, qui prend une forme différente selon que $\sigma_1 = \sigma_2$ ou non

Variances connues: test Z

- μ_1 et μ_2 sont tous deux inconnus
- σ_1 **et** σ_2 **sont connus**
- Statistique de test: $T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$
- Sous $H_0 : \mu_1 = \mu_2$, T suit une loi $\mathcal{N}(0, 1)$

Test bilatéral $H_0 : \mu_1 = \mu_2$ contre $H_1 : \mu_1 \neq \mu_2$

- Test de niveau α: rejeter de H_0 si $|t| \geq u$,

où u est le quantile d'ordre $1 - \alpha/2$ de la loi $\mathcal{N}(0, 1)$

Exemple: si $\alpha = 0.05$, alors $u = 1.959964$

- p-value du test: $p = \mathrm{P}(|\mathcal{N}(0, 1)| \geq |t|)$

Variances identiques: test de Student

- μ_1 et μ_2 sont tous deux inconnus
- $\sigma_1 = \sigma_2$, notée σ, où σ **est inconnu**
- Statistique de test: $T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S^2}{n_1} + \frac{S^2}{n_2}}}$, où $S^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$
- Sous $H_0 : \mu_1 = \mu_2$, T suit une loi de Student à $n1 + n2 - 2$ degrés de liberté ($t_{n_1+n_2-2}$)

Test bilatéral $H_0 : \mu_1 = \mu_2$ contre $H_1 : \mu_1 \neq \mu_2$

- Test de niveau α: rejeter de H_0 si $|t| \geq u$,

où u est le quantile d'ordre $1 - \alpha/2$ de la loi $t_{n_1+n_2-2}$

Exemple: si $\alpha = 0.05$, $n_1 = 20$ et $n_2 = 30$, alors $u = 2.0106348$

- p-value du test: $p = \mathrm{P}(|t_{n_1+n_2-2}| \geq |t|)$

Variances différentes: test de (Student-)Welch

- μ_1 et μ_2 sont tous deux inconnus
- $\sigma_1 \neq \sigma_2$ **et sont tous deux inconnus**
- Statistique de test: $T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$
- Sous $H_0 : \mu_1 = \mu_2$, T suit (approximativement) une loi de Student t_k, dont les degrés de liberté sont calculables en fonction de S_1, S_2, n_1, n_2 (formule de Welch-Satterthwaite)

Test bilatéral $H_0 : \mu_2 = \mu_1$ contre $H_1 : \mu_2 \neq \mu_1$

- Test de niveau α: rejeter de H_0 si $|t| \geq u$,

où u est le quantile d'ordre $1 - \alpha/2$ de la loi t_k

- p-value du test: $p = \mathrm{P}(|t_k| \geq |t|)$

Exemple: étude de l'effet d'un somnifère

Nombre d'heures de sommeil gagnées par 10 patients après la prise d'un somnifère

	SOMNIFÈRE 1	SOMNIFÈRE 2	DIFFÉRENCE ('1'-'2')
1	0.70	1.90	-1.20
2	-1.60	0.80	-2.40
3	-0.20	1.10	-1.30
4	-1.20	0.10	-1.30
5	-0.10	-0.10	0.00
6	3.40	4.40	-1.00
7	3.70	5.50	-1.80
8	0.80	1.60	-0.80
9	0.00	4.60	-4.60
10	2.00	3.40	-1.40

Cushny, A. R. and Peebles, A. R. (1905) The action of optical isomers: II hyoscines. The Journal of Physiology 32, 501–510.
Student (1908) The probable error of the mean. Biometrika, 6, 20.

Test d'homogénéité des variances de Fisher

Pour savoir si on peut considérer que $\sigma_1 = \sigma_2$, on effectue un **test de Fisher**

```
var.test(x, y)
```

```
##
##  F test to compare two variances
##
## data:  x and y
## F = 0.7983, num df = 9, denom df = 9, p-value = 0.7427
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.198297 3.214123
## sample estimates:
## ratio of variances
##          0.7983426
```

Ce test sera expliqué plus en détails en "Démarche Statistique II" dans le cours sur le modèle linéaire. Pour le moment nous nous contenterons de l'appliquer.

Test de Student sur échantillons indépendants

Mise en place du test

- H_0: "$\mu_1 = \mu_2$" contre H_1: "$\mu_1 \neq \mu_2$"
- Niveau choisi: $\alpha = 0.05$
- Présupposé: variances identiques (résultat du test de Fisher page précédente)
- Statistique de test: $T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}}$
- **Sous** H_0, $T \sim t(n_1 + n_2 - 2)$
- Forme du test: rejet de $H_0 \Leftrightarrow |T| \geq u$
- Seuil de rejet: $u = 2.100922$

Réalisation du test

- Données: x =0.7, -1.6, -0.2, -1.2, -0.1, 3.4, 3.7, 0.8, 0, 2; y=1.9, 0.8, 1.1, 0.1, -0.1, 4.4, 5.5, 1.6, 4.6, 3.4
- Statistique de test: -1.8608135
- p-value: 0.0791867

Conclusion

Au vu de cette expérience, au niveau 0.05, on ne peut rejeter l'hypothèse selon laquelle les deux somnifères ont la même efficacité

Test de Student sur échantillons indépendants

Pilote automatique

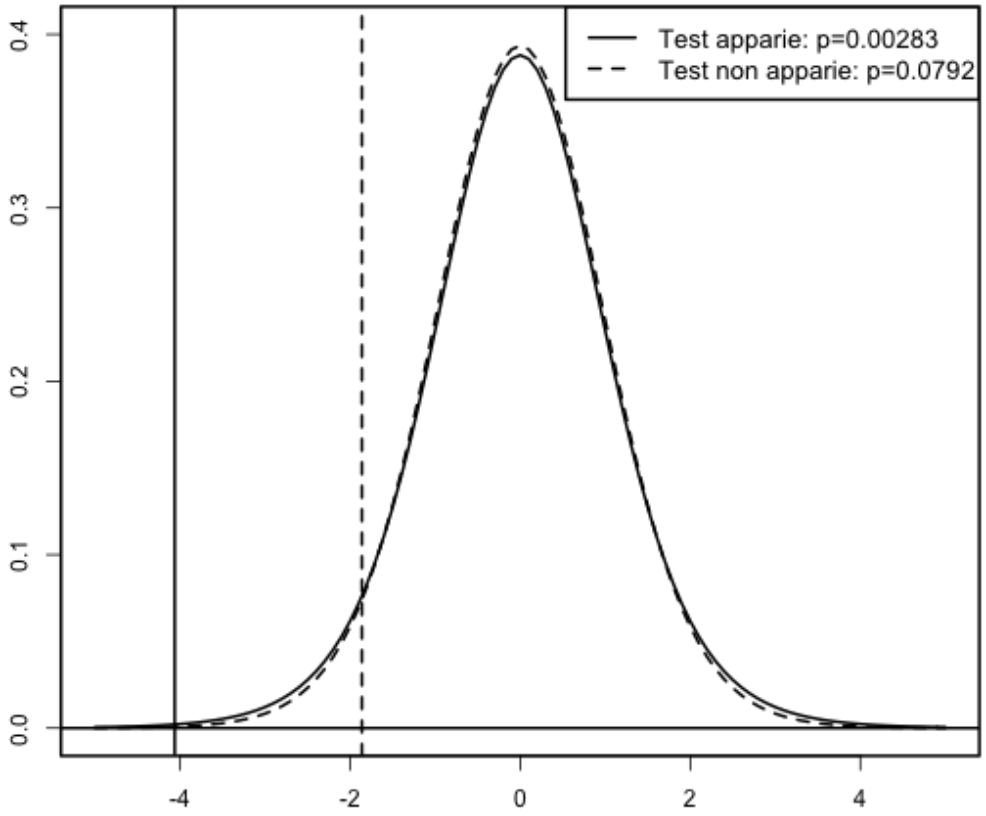
```
t.test(x, y, var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  x and y
## t = -1.8608, df = 18, p-value = 0.07919
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.363874  0.203874
## sample estimates:
## mean of x mean of y
##      0.75      2.33
```

Remarque: test sur échantillons appariés

Cas des données "somnifères"

- Les échantillons sont en fait appariés
- On **a le droit** de faire comme si les échantillons n'étaient pas appariés, et d'utiliser le test de Student sur échantillons indépendants.
- Le graphique ci-contre montre que **la prise en compte de l'appariement permet de mettre en évidence une différence significative (au seuil 5%)**



Test de Wilcoxon

Test de Wilcoxon

Motivation

Les tests de Student et Welch ci-dessus ne sont applicables que quand les échantillons $(X_1, \dots X_{n_1})$ et $(Y_1, \dots Y_{n_2})$ sont chacun iid de loi normale et indépendants entre eux.

Le test de Wilcoxon ne fait **pas d'hypothèse de normalité** (mais fait l'hypothèse d'indépendance !) Il suppose simplement qu'**ordonner les X_i et Y_j** a un sens.

Idée: tester l'égalité des médianes des deux échantillons

Remarque: ce test permet aussi de tester l'égalité des deux distributions

Mise en oeuvre

La statistique repose sur la *somme des rangs des X_i parmi l'ensemble des $n_1 + n_2$ valeurs*

Propriété

Sous H_0 (médiantes identiques), la loi de cette somme *ne dépend que de n_1 et n_2*

Test de Wilcoxon: application

Pilote automatique (test exact)

Comme pour les échantillons appariés: utiliser la fonction `wilcox.exact` du package `exactRankTests`

```
library(exactRankTests)
wilcox.exact(x, y, paired=FALSE)
```

```
##
##  Exact Wilcoxon rank sum test
##
## data:  x and y
## W = 25.5, p-value = 0.06582
## alternative hypothesis: true mu is not equal to 0
```

Remarque sur l'exemple "somnifères": on constate comme pour le test de Student que la prise en compte de l'appariement permettait de mettre en évidence une différence significative au seuil 5%