Démarche Statistique 1

Tests usuels sur échantillons appariés

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

Echantillons appariés

Définition

Deux séries d'observations sur les mêmes individus, mais dans des conditions différentes

n couples de variables *quantitatives* $(X_1, Y_1), (X_2, Y_2), \dots (X_n, Y_n)$ telles que les (X_i, Y_i) sont des variables indépendantes, identiquement distribuées. On note (μ_1, μ_2) la moyenne de (X, Y).

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

Exemples

- Efficacité de deux traitements laser pour la rétinopathie diabétique: un traitement par oeil; mesure de l'acuité visuelle un certain temps après traitement
- Effet secondaire d'un médicament contre le rhume: on se demande si le médicament n'aurait pas pour effet d'augmenter la tension artérielle. On mesure la tension de chaque patient; le patient prend le médicament; après un certain temps, on reprend la tension du patient

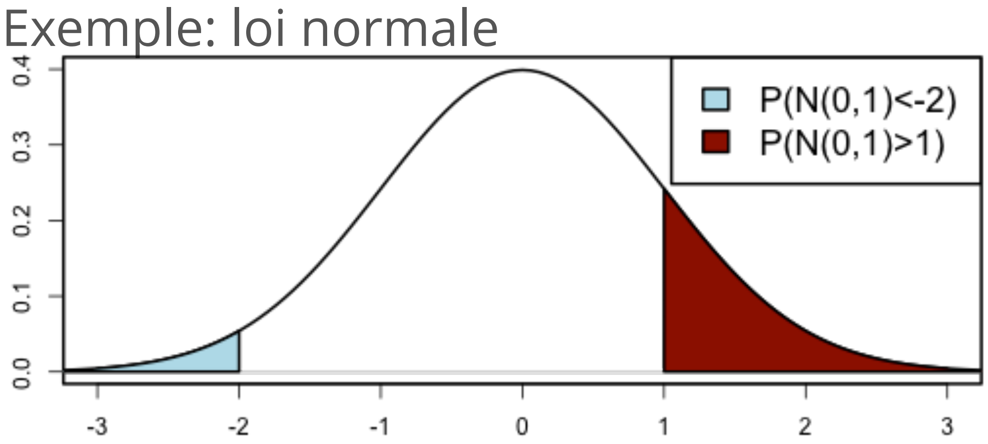
Plan

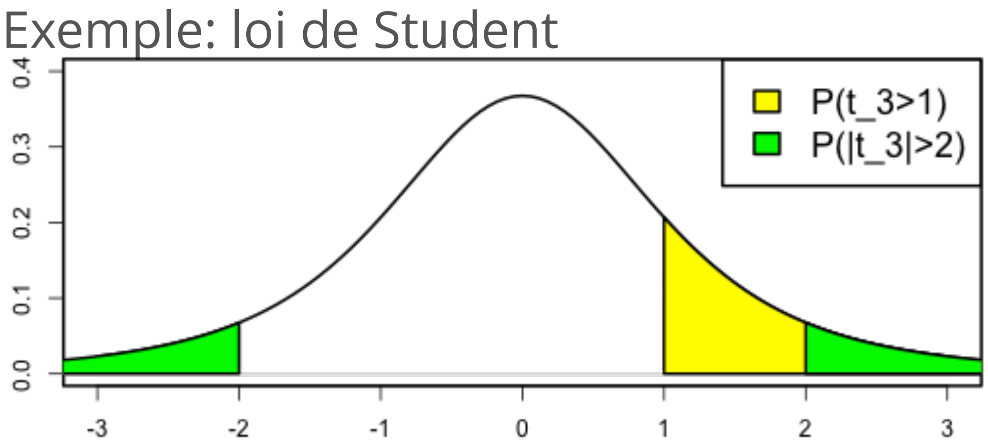
Notations

- On écrira que des variables sont **i.i.d.** (pour indépendantes, et identiquement distribuées) si elles sont indépendantes et suivent toutes la même loi de probabilité
- On utilisera des **lettres majuscules** pour désigner les **variables aléatoires** et des **lettres minuscules** pour désigner leurs **réalisations** (qui sont des valeurs numériques)
  Exemple:
  - $\bar{X} = \sum_{i=1}^{n} X_i$ désigne la moyenne empirique de l'échantillon $X_1, \ldots X_n$
  - Pour un échantillon donné, si les réalisations de ces variables sont $x_1 = 2$, $x_2 = 0$, $x_3 = 1$, alors la **valeur** de la moyenne empirique de l'échantillon est $\bar{x} = 1$

Notations (suite)

- On notera $\mathrm{P}(\mathcal{N}(0, 1) \geq x)$ la probabilité qu'une variable aléatoire de loi $\mathcal{N}(0, 1)$ dépasse la valeur x, c'est-à-dire l'aire sous la courbe de densité à droite de x.
- En `R`, on a $\mathrm{P}(\mathcal{N}(0, 1) \geq x) =$ `1-pnorm(x)`
- On utilisera de même les notations:
  - $\mathrm{P}(\mathcal{N}(0, 1) \leq x) =$ `1-pnorm(x)`
  - $\mathrm{P}(|\mathcal{N}(0, 1)| \geq x) =$ `2*(1-pnorm(x))`
- On utilisera également cette notation pour d'autres lois (en particulier la loi de Student à k degrés de liberté, notée t_k).





Cette notation n'est pas ambigüe car il s'agit de quantités intrinsèques à la loi de probabilité

1. Test de Student

Données appariées: méthode des couples

Principe: travailler sur les différences entre observations

On note $D_i = Y_i - X_i$ pour tout i.

On suppose que $D_1, D_2, \ldots D_n$ sont indépendantes, identiquement distribuées de loi $\mathcal{N}(\mu_2 - \mu_1, \sigma^2)$.

Pour tester $H_0 : \mu_1 = \mu_2$ contre $H_1 : \mu_1 \neq \mu_2$ (ou $H_1 : \mu_1 < \mu_2$), on peut utiliser le test de comparaison à la valeur de référence 0.

$$\bar{D} = \frac{\sum_{i=1}^{n} D_i}{n}$$

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (D_i - \bar{D})^2$$

On présente ici le cas où σ est inconnu car c'est le plus réaliste. On utilse un test de Student.

Test de Student sur données appariées

- Les différences $D_1, \ldots D_n$ sont supposées i.i.d. de loi $\mathcal{N}(\mu_2 - \mu_1, \sigma^2)$
- μ_1, μ_2 **et** σ **sont inconnus**
- Statistique de test: $T = \frac{\bar{D}}{S/\sqrt{n}}$
- Sous $H_0 : \mu_1 = \mu_2$, T suit une loi de Student à $n-1$ degrés de liberté

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 de Student à $n-1$ degrés de liberté

Exemple: si $\alpha = 0.05$ et $n = 20$, alors $u = 2.0930241$

- p-value du test: $p = \mathrm{P}_{H_0}(|T| \geq |t|)$

Test de Student sur données appariées

- Les différences $D_1, \ldots D_n$ sont supposées i.i.d. de loi $\mathcal{N}(\mu_1 - \mu_2, \sigma^2)$
- **μ_1, μ_2 et σ sont inconnus**
- Statistique de test: $T = \frac{\bar{D}}{S/\sqrt{n}}$
- Sous $H_0 : \mu_1 = \mu_2$, T suit une loi de Student à $n-1$ degrés de liberté: t_{n-1}

Test unilatéral à gauche

- $H_0 : \mu_2 = \mu_1$ contre $H_1 : \mu_1 < \mu_2$
- Test de niveau α: rejet de H_0 si $t \leq u$,

où u est le quantile d'ordre α de la loi t_{n-1}

Si $\alpha = 0.05$, alors $u = -1.7291328$

- p-value du test: $p = \mathrm{P}_{H_0}(T \leq t)$

Test unilatéral à droite

- $H_0 : \mu_2 = \mu_1$ contre $H_1 : \mu_1 > \mu_2$
- Test de niveau α: rejet de H_0 si $t \geq u$,

où u est le quantile d'ordre $1-\alpha$ de t_{n-1}

Si $\alpha = 0.05$, alors $u = 1.7291328$

- p-value du test: $p = \mathrm{P}_{H_0}(T \geq t)$

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 de Student sur données appariées

Mise en place du test

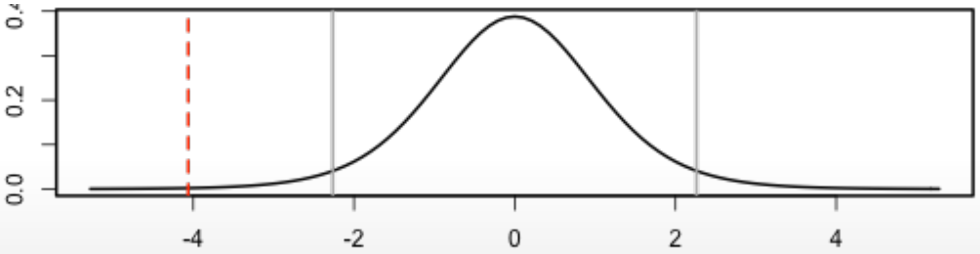
- Test de H_0: "$\mu_2 = \mu_1$" contre H_1: "$\mu_1 \neq \mu_2$"
- Niveau choisi: $\alpha = 0.05$
- Statistique de test: $T = \frac{\bar{D}}{S/\sqrt{n}}$ pour $n = 10$
- **Sous** H_0, T suit une loi de Student $t(n-1)$, **connue**
- Forme du test: rejet de $H_0 \Leftrightarrow |T| \geq u$
- Seuil de rejet: $u = 2.2621572$

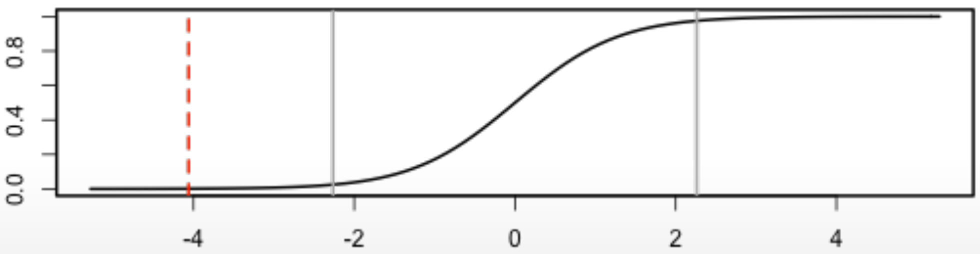
Réalisation du test

- Données: -1.2, -2.4, -1.3, -1.3, 0, -1, -1.8, -0.8, -4.6, -1.4
- Statistique de test: -4.0621277
- p-value: 0.0028329

Conclusion

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





Test de Student sur données appariées

Statistique de test

```
d <- x-y
S <- sd(d)
n <- length(d)
(stat <- mean(d)/S*sqrt(n))
```

```
## [1] -4.062128
```

Calcul du quantile

```
alpha <- 0.05
(u <- qt(1-alpha/2, df=n-1))
```

```
## [1] 2.262157
```

Décision

```
ifelse(abs(stat)>=u, "rejet", "non rejet")
```

```
## [1] "rejet"
```

Calcul de la p-valeur

```
(p <- 2*(1-pt(abs(stat), df=n-1)))
```

```
## [1] 0.00283289
```

Test de Student sur données appariées

Pilote automatique

```r
t.test(x, y, paired=TRUE)
```

```
##
##  Paired t-test
##
## data:  x and y
## t = -4.0621, df = 9, p-value = 0.002833
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.4598858 -0.7001142
## sample estimates:
## mean of the differences
##                   -1.58
```

2. Test du signe

Test du signe

Motivation

Le test de Student n'est applicable que quand les D_i sont indépendantes et de loi normale.

Le test du signe ne fait **pas d'hypothèse de normalité** (mais fait l'hypothèse d'indépendance !)

Idée: tester la symétrie de la distribution de D par rapport à 0

Mise en oeuvre

- Compter le nombre T d'individus dont le D_i est positif

Propriété

Sous l'hypothèse de symétrie de la loi de D (c'est-à-dire sous l'hypothèse que la médiane de D est nulle), la loi de T est $Bin(n, 1/2)$

3. Test des rangs signés de Wilcoxon

Test de Wilcoxon

Motivation

Le test de Student n'est applicable que quand les D_i sont indépendantes et de loi normale.

Le test de Wilcoxon ne fait **pas d'hypothèse de normalité** (mais fait l'hypothèse d'indépendance !) Il suppose simplement qu'**ordonner les** D_i a un sens (variables ordinales)

Idée: tester la symétrie de la distribution de D par rapport à 0

Mise en oeuvre

- Calculer les rangs R_i des valeurs absolues des différences D_i et le signe s_i de la différence
- Calculer la *somme des rangs des individus dont le D_i est positif*: $T = \sum_{\{D_i>0\}} R_i$

Propriété

Sous l'hypothèse de symétrie de la loi de D (c'est-à-dire sous l'hypothèse que la médiane de D est nulle), la loi de ces variables aléatoires ne dépend que de n

Exemple: somnifères

Mise en place du test

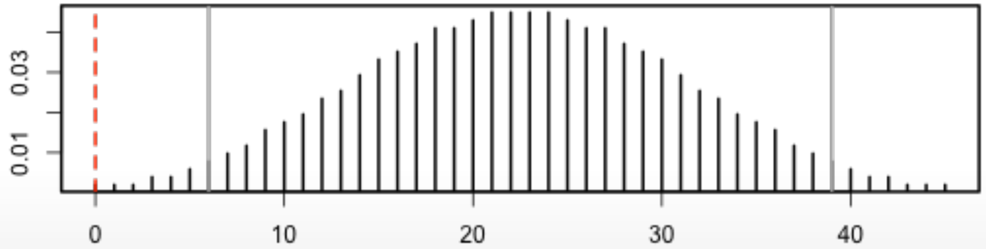
- Test de H_0: "$\mu_1 = \mu_2$" contre H_1: "$\mu_1 \neq \mu_2$"
- Niveau choisi: $\alpha = 0.05$
- Statistique de test: $T = \sum_{\{D_i>0\}} R_i$
- **Sous** H_0, T suit une loi de Wilcoxon(n), **connue**
- Forme: rejet de $H_0 \Leftrightarrow T \leq u_g$ ou $T \geq u_d$
- Seuils de rejet: $u_g = 6, u_d = 39$

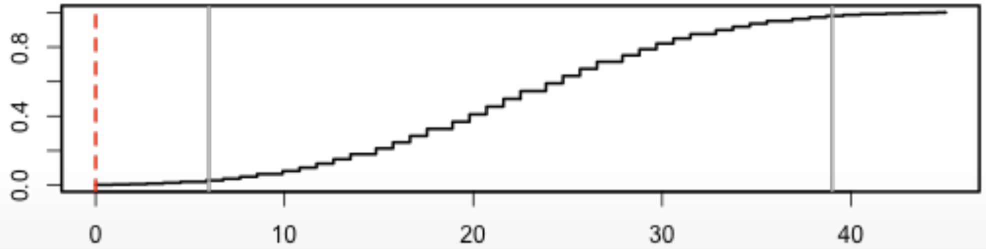
Réalisation du test

- Données: -1.2, -2.4, -1.3, -1.3, -1, -1.8, -0.8, -4.6, -1.4
- Statistique de test: 0
- p-value: 0.0039062

Conclusion

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





Test de Wilcoxon sur données appariées

Statistique de test

```
d <- x-y; d <- d[d != 0]
n <- length(d)
r <- rank(abs(d))
(stat <- sum(r[d > 0]))
```

```
## [1] 0
```

Calcul des quantiles

```
alpha <- 0.05
(u <- qsignrank(c(alpha/2, 1-alpha/2), n))
```

```
## [1]  6 39
```

Décision

```
ifelse(stat>=u[1] || stat<=u[2], "rejet", "non rejet")
```

```
## [1] "rejet"
```

Calcul de la p-valeur

NB: la loi de la statistique sous H_0 n'est pas symétrique ⇒ calcul de la p-valeur adapté

```
pg <- psignrank(stat, n)
(p <-2*min(pg, 1-pg))
```

```
## [1] 0.00390625
```

Test de Wilcoxon sur données appariées

Pilote automatique ?

```
suppressWarnings(wilcox.test(x, y, paired=TRUE))
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  x and y
## V = 0, p-value = 0.009091
## alternative hypothesis: true location shift is not equal to 0
```

- Noter les "*`Warning`*" à cause de la présence de 0 et d'ex-aequo (ties) et le "*`with continuity correction`*"
- Utiliser plutôt la fonction `wilcox.exact` du package `exactRankTests`

Test de Wilcoxon sur données appariées

Pilote automatique: test exact

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

```
##
##  Exact Wilcoxon signed rank test
##
## data:  x and y
## V = 0, p-value = 0.003906
## alternative hypothesis: true mu is not equal to 0
```

Ce test utilise la loi exacte de la statistique de test: **c'est celui qu'il faut utiliser** !

Approximation gaussienne

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

```
##
##  Asymptotic Wilcoxon signed rank test
##
## data:  x and y
## V = 0, p-value = 0.007632
## alternative hypothesis: true mu is not equal to 0
```

Ce test utilise une approximation de la loi de la statistique de test par une loi normale, qui n'est valable que pour des échantillons de taille suffisante (ce n'est pas le cas ici car $n = 9$)