Cours Statistiques et Analyse de données

Un bout du cours 2 que vous auriez pu avoir la dernière fois

MASTER INFORMATIQUE - PARCOURS SIIA

G. Coppin

2020-2021

Objectifs du cours

Objectifs :

1. Maîtriser les principaux tests statistiques

Méthodologie générale des tests (I)

- Enoncer l'hypothèse à tester. $H_0 : \theta = \theta_0$.
- Se donner un risque d'erreur α
- Tirer un échantillon de la population - attention, méthodologie à suivre (cf. cours 1)
- Calculer un estimateur du paramètre - par exemple $\overline{X}$ pour estimer une moyenne
- Etudier l'écart entre $\widehat{\theta}$ et θ_0. Si cet écart est grand, l'hypothèse est rejetée.
- Tirer une conclusion : On peut donc avoir rejeté H_0 ou ne pas avoir pu le faire. Alors deux types d'erreur possibles
  - rejeter même lorsqu'elle est vraie (risque α)
  - accepter alors qu'elle est fausse - difficile à évaluer ("degré de fausseté" de l'hypothèse H_0)

Méthodologie générale des tests (II)

Une conserverie met sur le marché des boites de petits pois dont l'étiquette mentionne 400g. Le directeur de la production veut vérifier que le poids est bien respecté.

- $H_0 : \mu = 400g$
- soit H_0 est rejetée pour un poids trop lourd ou trop léger. Si boites ok, coût inutile. Risque α
- soit H_0 acceptée à tort mais en réalité possibilité que le marché soit inondé de boites trop lourdes (pertes) ou trop légères (arnaque). Probabilité de ce type d'erreur non contrôlée.

Test d'hypothèse sur une proportion (I)

X de loi $B(n,p)$ avec n connu et p inconnu.

- Hypothèse nulle : $H_0 : p = p_0$ avec p_0 donné
- Risque α fixé
- Si H_0 vraie, $\widehat{p} = X/n$ l'estimateur de p est de loi $N(p_0, \frac{p_0 q_0}{n})$.
  Donc si H_0 vraie, on a $Z = \frac{\sqrt{n}(\widehat{p}-p_0)}{\sqrt{p_0 q_0}}$ sera de loi $N(0,1)$

Test d'hypothèse sur une proportion (II)

L'hypothèse nulle est rejetée si Z est trop grand ou trop petit. soit si Z est en dehors de l'intervalle $[-c_\alpha, +c_\alpha]$. Autrement dit

- H_0 est rejetée si $\mid Z \mid > c_\alpha$ ou aussi $\mid \widehat{p} - p_0 \mid > \frac{\sqrt{p_0 q_0}}{\sqrt{n}}$
- H_0 est acceptée si $\mid Z \mid \leq c_\alpha$ ou aussi $\mid \widehat{p} - p_0 \mid \leq \frac{\sqrt{p_0 q_0}}{\sqrt{n}}$

Test d'hypothèse sur une proportion (III)

On fait l'hypothèse que 25% des gens sont gauchers. Avec $\alpha = 10\%$, on trouve 18 gauchers sur 120 personnes.

- $p_0 = 0{,}25$ et $\widehat{p} = 0{,}15$
- $c_\alpha = 1,645$. Donc $c_\alpha \sqrt{p_0 q_0}/\sqrt{n} = 0,065$

Puisque $|\widehat{p} - p_0| = 0,25 - 0,15 = 0,10 > 0,065$ (on a obtenu une valeur "trop extrême"), on rejette l'hypothèse nulle. Il y certainement moins de gauchers.

En R, ça donne ...

```
> prop.test(18, 120, 0.25)
1-sample proportions test with continuity correction
data : 18 out of 120, null probability 0.25
X-squared = 5.8778, df = 1, p-value = 0.01533
alternative hypothesis : true p is not equal to 0.25
95 percent confidence interval :
0.09369541 0.22939185
sample estimates :
p
0.15
```

mais aussi

```
> binom.test(18, 120, 0.25)
Exact binomial test
data : 18 and 120
number of successes = 18, number of trials = 120, p-value = 0.01102
alternative hypothesis : true probability of success is not equal to 0.25
95 percent confidence interval :
0.09138957 0.22666714
sample estimates :
probability of success
0.15
```

Test d'égalité sur deux proportions (I)

Dans un étude américaine qui portait sur le taux de mortalité de 92 patients sérieusement cardiaques, 53 de ces patients avaient un animal familier et parmi ces 53 patients, 3 sont morts dans l'année. Parmi les 39 qui n'avaient pas d'animal, 11 sont morts dans l'année. Peut-on dire que les deux groupes avaient la même chance d'y passer ?
Autrement dit, $\widehat{p} = 0,057$ et $\widehat{p} = 0,282$ ont ils un écart significatif vue la taille de l'échantillon ?

Test d'égalité sur deux proportions (II)

- on a deux lois $B(n_x, p_x)$ et $B(n_y, p_y)$
- on veut tester l'hypothèse $H_0 : p_x = p_y$
- $\widehat{p}_x$ est de loi $N(p_x, \sigma^2_{\widehat{p}_x})$ avec $\sigma^2_{\widehat{p}_x} = p_x q_x / n_x$
- idem en y
- les deux échantillons sont a priori indépendants

Test d'égalité sur deux proportions (III)

Indépendants, donc on peut soustraire et donc

$$\frac{\widehat{p}_x - \widehat{p}_y - (p_x - p_y)}{\sqrt{\sigma^2_{\widehat{p}_x} + \sigma^2_{\widehat{p}_y}}}$$

est de loi $N(0,1)$.

Si H_0 est vraie, on a $p_x = p_y$ et donc $Z = \frac{\widehat{p}_x - \widehat{p}_y}{\sqrt{\sigma^2_{\widehat{p}_x} + \sigma^2_{\widehat{p}_y}}}$ est approx. de loi $N(0,1)$. On peut appliquer un test de risque α comme précédemment.

Test d'égalité sur deux proportions (IV) - le retour de Docteur House

$$Z = \frac{\widehat{p}_x - \widehat{p}_y}{\sqrt{\widehat{\sigma^2_{\widehat{p}_x}} + \widehat{\sigma^2_{\widehat{p}_y}}}}$$

$$Z = \frac{0,057 - 0,282}{\sqrt{0,00101 + 0,00519}} = -2,86$$

Test d'égalité sur deux proportions (IV) - le retour de Docteur House

$$Z = \frac{\widehat{p}_x - \widehat{p}_y}{\sqrt{\widehat{\sigma^2_{\widehat{p}_x}} + \widehat{\sigma^2_{\widehat{p}_y}}}}$$

$$Z = \frac{0,057 - 0,282}{\sqrt{0,00101 + 0,00519}} = -2,86$$

Même avec $\alpha = 1\%$ (et $c_\alpha = 2,576$), on a quand même une valeur trop extrême. Donc on rejette H_0. Au tour de Docteur House d'interpréter le résultat ...

Et en R ?

```
> prop.test(c(3,11), c(53,39))
2-sample test for equality of proportions with continuity correction
data : c(3, 11) out of c(53, 39)
X-squared = 7.1899, df = 1, p-value = 0.007331
alternative hypothesis : two.sided
95 percent confidence interval :
-0.4020273 -0.0488677
sample estimates :
prop 1 prop 2
0.05660377 0.28205128
```

Test d'hypothèse sur une moyenne (I)

On veut tester une moyenne, soit $H_0 : \mu = \mu_0$

- On sait que l'estimateur naturel de μ est $\overline{X}$ et que $\frac{\overline{X}-\mu}{\frac{\widehat{\sigma}}{\sqrt{n}}}$ est pratiquement de loi $N(0,1)$.
- Quand les lois sont normales, $T = \frac{\overline{X}-\mu}{\frac{\widehat{\sigma}}{\sqrt{n}}}$ est exactement de loi de Student à $n-1$ degrés de liberté. On va s'aligner sur cette configuration.

Test d'hypothèse sur une moyenne (II)

Dans ce cas

- H_0 est rejetée si $\mid T \mid > c_\alpha$, soit $\mid \overline{X} - \mu_0 \mid > \frac{c_\alpha \widehat{\sigma}}{\sqrt{n}}$
- H_0 est acceptée si $\mid T \mid \leq c_\alpha$ soit $\mid \overline{X} - \mu_0 \mid \leq \frac{c_\alpha \widehat{\sigma}}{\sqrt{n}}$

Test d'hypothèse sur une moyenne (II) - exemple

On fait l'hypothèse que le temps moyen de sommeil est de 7,7 heures. Un labo vend un somnifère miracle et obtient les résultats suivants sur un échantillon de taille 10.

7,8	8,3	7,2	9,1	8,4	6,8	7,3	7,7	8,9	9,2

Avec $\alpha = 5\%$, on teste $H_0 : \mu = \mu_0 = 7,7$

Test d'hypothèse sur une moyenne (II) - exemple

- $n = 10, \sum X_i = 80,7$ et $\sum X_i^2 = 657,61$ ce qui donne $\overline{X} = 8,07$ et $\widehat{\sigma} = 0,8407$
- $T = \frac{\overline{X} - \mu}{\frac{\widehat{\sigma}}{\sqrt{n}}} = \frac{\sqrt{10}(8,07-7,7)}{0,8407} = 1,392$

Avec $\nu = 9$ et $\alpha = 5\%$, $c_\alpha = 2,262$. Donc $\mid T \mid < c_\alpha$ et on ne rejette pas l'hypothèse nulle. Escroc de laboratoire ! ...

Avec votre logiciel favori ...

```
> t.test(c(7.8, 8.3, 7.2, 9.1, 8.4, 6.8, 7.3, 7.7, 8.9, 9.2), mu = 7.7,
conf.level = 0.95)
One Sample t-test
data : c(7.8, 8.3, 7.2, 9.1, 8.4, 6.8, 7.3, 7.7, 8.9, 9.2)
t = 1.3917, df = 9, p-value = 0.1974
alternative hypothesis : true mean is not equal to 7.7
95 percent confidence interval :
7.468599 8.671401
sample estimates :
mean of x
8.07
```

Avec votre logiciel favori ...

```
> t.test(c(7.8, 8.3, 7.2, 9.1, 8.4, 6.8, 7.3, 7.7, 8.9, 9.2), mu = 7.7,
conf.level = 0.95)
One Sample t-test
data : c(7.8, 8.3, 7.2, 9.1, 8.4, 6.8, 7.3, 7.7, 8.9, 9.2)
t = 1.3917, df = 9, p-value = 0.1974
alternative hypothesis : true mean is not equal to 7.7
90 percent confidence interval :
7.582662 8.557338
sample estimates :
mean of x
8.07
```

Test d'hypothèse sur une moyenne (II) - le cas Marcel Schblurb

Vous êtes sur le point d'être embauché(e) dans la société Macheprot. Marcel-Benoit Schblurb, de la promo précédente, vous indique que le salaire moyen des embauchés est de 35 keuros. Apres petite enquête, vous obtenez les données suivantes :

Emb1	Emb2	Emb3	Emb4	Emb5
34.5	36	35.2	33	34.3

A vous de jouer ...

Test d'égalité de deux moyennes (I)

On compare deux échantillons $X_1, .., X_n$ et $Y_1, .., Y_p$ provenant de deux populations et on veut tester l'hypothèse

$$H_0 : \mu_X = \mu_Y$$

Si n et p sont assez grands, on sait que $\overline{X}$ et $\overline{Y}$ suivent respectivement des lois $N(\mu_X, \frac{\sigma_X^2}{n})$ et $N(\mu_Y, \frac{\sigma_Y^2}{p})$. Si on estime que les populations et donc les variables sont indépendantes, on sait que $\overline{X} - \overline{Y}$ est de loi $N(\mu_X - \mu_Y, \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{p})$.

Test d'égalité de deux moyennes (II)

Donc $\frac{\overline{X}-\overline{Y}-(\mu_X-\mu_Y)}{\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{p}}}$ est approximativement de loi $N(0,1)$. Si H_0 est vraie,

$$\frac{\overline{X}-\overline{Y}}{\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{p}}}$$

est approximativement de loi $N(0,1)$ et, en approximant les variances inconnues par les mesures issues des observations,

$$Z=\frac{\overline{X}-\overline{Y}}{\sqrt{\frac{\widehat{\sigma}_X^2}{n}+\frac{\widehat{\sigma}_Y^2}{p}}}$$

l'est aussi. H_0 sera rejetée pour $\mid Z \mid > c_\alpha$ trop extrême.

Test d'égalité de deux moyennes (III) - retour sur le jeunisme de la filière 3

$\overline{X} = 24,6$, $\overline{Y} = 25,1$, $n = 35$, $p = 22$, $\sigma_X = 0,4$, $\sigma_Y = 0,3$. Alors :

$$Z = \frac{25,1 - 24,6}{\sqrt{\frac{0,4}{35} + \frac{0,3}{22}}} = 3.16$$

Pour $\alpha = 5\%$, Z est trop extrême pour que H_0 soit acceptée, et donc oui, il y a une différence et cette différence n'est pas expliquée par le hasard...

Test d'égalité de deux moyennes (III) - variances égales

Lorsque les variances sont supposées égales (hypothèse très courante), la formule se simplifie et la loi $\frac{\overline{X}-\overline{Y}}{\sigma\sqrt{\frac{1}{n}+\frac{1}{p}}}$ est $N(0,1)$. On estime $\widehat{\sigma}$ à partir de $\widehat{\sigma}_X$ et $\widehat{\sigma}_Y$ à l'aide de l'estimateur sans biais :

$$\widehat{\sigma} = \frac{(n-1)\widehat{\sigma}_X^2 + (p-1)\widehat{\sigma}_Y^2}{n+p-2}$$

et

$$T = \frac{\overline{X}-\overline{Y}}{\widehat{\sigma}\sqrt{\frac{1}{n}+\frac{1}{p}}}$$

est exactement une loi de Student à $n+p-2$ degrés de liberté. L'hypothèse H_0 sera acceptée si $\mid T \mid \leq c_\alpha$.

RRRRR - cf. le film

```
> t.test(c(2,3,5,6,3), c(2,4,5,1,1), var.equal = FALSE)
Welch Two Sample t-test
data : c(2, 3, 5, 6, 3) and c(2, 4, 5, 1, 1)
t = 1.0954, df = 7.921, p-value = 0.3055
alternative hypothesis : true difference in means is not equal to 0
95 percent confidence interval :
-1.330505 3.730505
sample estimates :
mean of x mean of y
3.8 2.6

> t.test(c(2,3,5,6,3), c(2,4,5,1,1), var.equal = TRUE)
Two Sample t-test
data : c(2, 3, 5, 6, 3) and c(2, 4, 5, 1, 1)
t = 1.0954, df = 8, p-value = 0.3052
alternative hypothesis : true difference in means is not equal to 0
95 percent confidence interval :
-1.326101 3.726101
sample estimates :
mean of x mean of y
3.8 2.6
```

Test d'égalité de deux moyennes (IV) - données appariées

Si on veut mesurer l'effet d'un traitement sur une population, on a les mêmes individus lors des deux tests ($n = p$). On ne peut plus supposer l'indépendance et rester dans le cadre précédent. Il suffit de prendre la variable $W_i = X_i - Y_i$ et l'hypothèse nulle devient $H_0 : \mu_W = 0$. On peut se ramener au test de valeur de moyenne, avec une variable de Student à $n - 1$ degrés de liberté.
Si on ne prend pas ces précautions, on surestime la variance $(\sigma_X^2 + \sigma_Y^2)$ et on peut aboutir à des acceptations biaisées de l'hypothèse nulle.

Test d'égalité de deux moyennes (V) - exemple

Quelques E/C de Lussi ont enfin décidé de perdre du poids et sont résolus à ne plus manger de pain avec leurs nouilles. Leurs poids avant et après le régime sont les suivants :

Sujet	1	2	3	4	5	6
Avant	64	54	73	59	64	68
Après	61	54	71	58	61	66

Ont-ils vraiment bien fait de se priver ($\alpha = 5\%$) ?

Test d'égalité de deux moyennes (V) - exemple

Quelques E/C de Lussi ont enfin décidé de perdre du poids et sont résolus à ne plus manger de pain avec leurs nouilles. Leurs poids avant et après le régime sont les suivants :

Sujet	1	2	3	4	5	6
Avant	64	54	73	59	64	68
Après	61	54	71	58	61	66

Ont-ils vraiment bien fait de se priver ($\alpha = 5\%$) ?
$H_0 : \mu_{av} = \mu_{ap}$, $\overline{X} - \overline{Y} = 1,833$, $\widehat{\sigma}^2 = 1,367$, $T = 3,84$, $\nu = 5$,
$c_\alpha = 2,571$ donc on peut estimer que le régime marche.

avec la lettre qui suit Q

```
> t.test(c(64, 54, 73, 59, 64, 68), c(61, 54, 71, 58, 61, 66), var.equal = TRUE, paired = FALSE)
Two Sample t-test
data : c(64, 54, 73, 59, 64, 68) and c(61, 54, 71, 58, 61, 66)
t = 0.502, df = 10, p-value = 0.6266
alternative hypothesis : true difference in means is not equal to 0
95 percent confidence interval :
-6.304374 9.971041
sample estimates :
mean of x mean of y
63.66667 61.83333

> t.test(c(64, 54, 73, 59, 64, 68), c(61, 54, 71, 58, 61, 66), var.equal = TRUE, paired = TRUE)
Paired t-test
data : c(64, 54, 73, 59, 64, 68) and c(61, 54, 71, 58, 61, 66)
t = 3.8414, df = 5, p-value = 0.01211
alternative hypothesis : true difference in means is not equal to 0
95 percent confidence interval :
0.6064956 3.0601710 sample estimates :
mean of the differences
1.833333
```