Introduction aux tests statistiques

Antonio Falcó

Objectif

La logique des tests statistiques permet de formaliser la façon de tirer des conclusions à partir d'une expérience.

L'hypothèse nulle

Un test statistique se base sur les points suivants:

- Formulation d'une hypothèse. Traditionnellement, on appelle cette hypothèse l'hypothèse nulle et on la note H_0. (La raison de cette appellation apparaîtra plus clairement dans la suite.)
- Utilisation de la théorie statistique pour déterminer si les données soutiennent cette hypothèse H_0 ou non.
- Rejet de H_0 si les données ne la soutiennent pas.

Exemple

- Dans l'exemple de la piéce, on fait l'hypothèse qu'on a une pièce équilibrée:
$$H_0 : \Pr(\text{Pile}) = \Pr(\text{Face}) = 1/2.$$
- D'une façon mathématique: Soit $X \sim \mathcal{B}(p)$ avec $X(\omega) = 0$ si $\omega = \text{Pile}$ ou $X(\omega) = 1$ si $\omega = \text{Face}$ et
$$H_0 : p = 1/2.$$
- Question: A quel point l'observation d'un jet Pile sur un, de deux jets Pile sur deux,..., de six jets Pile sur six soutiennent cette hypothèse?

Objectif

- Dans ce qui suit, on va en fait calculer à quel point ces observations condamnent cette hypothèse.
- Pour ce faire on va calculer, sous l'hypothèse H_0, la probabilité que les observations s'éloignent au moins autant de H_0 que ce qui a été observé.
- Si cette probabilité est faible, on en conclura que soit H_0 n'est pas vraie, soit un événement rare a eu lieu.
- Ne croyant pas en la survenue d'un événement rare, on rejettera alors H_0.

Soit $X_1, \ldots, X_n$ un échantillon d'une population $X \sim \mathcal{B}(p)$. Alors

$$\widehat{p} := \frac{X_1 + X_2 + \cdots + X_n}{n} = \overline{X} \text{ est une approximation de } p$$

ou p est une variable aléatoire. Le TCL nous dit que si $n > 30$ alors

$$\widehat{p} := \frac{X_1 + X_2 + \cdots + X_n}{n} = \overline{X} \sim \mathcal{N}\left(p, \frac{\sqrt{p(1-p)}}{n}\right)$$

On sait que

$$Z = \frac{\widehat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim \mathcal{N}(0, 1),$$

et si on connaît la valeur du paramétre p on peux calculer, pour chaque valeur $0 < \alpha < 1$ fixée, le nombre x_α tel que

$$x_\alpha = \mathtt{qnorm}(1 - \alpha/2) \Leftrightarrow \Pr(-x_\alpha \leq Z \leq x_\alpha) = 1 - \alpha.$$

```
alpha ← 0.05
x.alpha ← qnorm(1−alpha/2)
x.alpha
```

```
[1] 1.959964
```

$$1.96 = \mathtt{qnorm}(0.95) \Leftrightarrow \Pr(-1.96 \leq Z \leq 1.96) = 0.95.$$

```
imcenfant ← read.csv2(
"~/Dropbox/Cursos/Biostatistique/Cours/Donnees/imcenfant.csv")
names(imcenfant)
```

```
[1] "SEXE"    "zep"     "poids"   "an"      "mois"    "taille"
```

```
table(imcenfant$SEXE)/length(imcenfant$SEXE)
```

```
        F         G
0.4671053 0.5328947
```

L'objectif est de répondre la question scientifique suivant:
Est-ce la probabilité d'être garçon est différent d'être fille dans la population?

Stratégie

On va supposé que c'est fausse, alors l'hypothèse $\Pr(F) = \Pr(G) = 0.5$ doit se verifé avec un haute niveaux de probabilité, disons $1-\alpha$ avec α petit (par exemple, $\alpha = 0.05$ alors $1-\alpha = 0.95$). En conséquence on suppose que l'hypothèse

$$H_0 : \Pr(F) = \Pr(G) = 0.5 \text{ est vrai avec le } 1-\alpha = 0.95 \text{ de probabilité,}$$

front l'hypothèse

$$H_1 : \Pr(F) \neq \Pr(G) = 0.5 \text{ est vrai avec le } \alpha = 0.05 \text{ de probabilité,}$$

On fait un tirage au sort dans la population de référence un utilisant un échantillon on peut calculer $\widehat{p}$:

```
taille.SEXE.F ← imcenfant$taille[imcenfant$SEXE=="F"]
taille ← imcenfant$taille
p.chapeau ← length(taille.SEXE.F)/length(taille)
p.chapeau
```

```
[1] 0.4671053
```

Si l'hypothèse H_0 est vrai avec une proba de 95%, on connait que $p = 0.5$ et comme $Z = \frac{\widehat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \sim \mathcal{N}(0,1)$ alors

$$\Pr\left(-1.96 \leq \frac{\widehat{p}-0.5}{\sqrt{\frac{0.5(1-0.5)}{n}}} \leq 1.96\right) = 1-\alpha = 0.95$$

est verifié.

On va calculer $Z = \frac{\widehat{p}-p}{\frac{\sqrt{p(1-p)}}{n}}$ sous la condition que l'hypothèse H_0 est vrai:

```
n ← length(imcenfant$SEXE)
n
```

```
[1] 152
```

```
Z ← (p.chapeau − 0.5)/(sqrt(0.5^2)/n)
Z
```

```
[1] -10
```

Comme $Z = -10$ est dehors de l'intervalle $[-1.96, 1.96]$ alors on arrive a une contradiction avec le fait que Z, sous H_0, est contenue dans l'intervalle $[-1.96, 1.96]$ avec le 95% de probabilité. En conséquence, on peut rejeter l'hypothèse H_0 avec un niveau $\alpha = 0.05$. La réponse à la question scientifique est positive.

La question scientifique est maintenant la suivante: En supposant que $\texttt{taille} = X \sim \mathcal{N}(\mu, \sigma)$ avec $\sigma = 4$ cm, est-ce la taille moyenne chez la population est différent de 100 cm? On sait

$$\texttt{mean(taille)} = \overline{X} = \frac{X_1 + \cdots + X_n}{n} \sim \mathcal{N}\left(\mu, \frac{\sigma = 4}{\sqrt{n}}\right)$$

ou $X_1, \ldots, X_n$ est un échantillon de $X \sim \mathcal{N}(\mu, \sigma)$:

Test d'hypothèse

$$H_0 : \mu = 100 \text{ cm},$$

avec le 95%

$$H_1 : \mu \neq 100 \text{ cm}.$$

avec 5%.

Si H_0 est vrai, comme

$$\Pr\left(-1.96 \leq \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq 1.96\right) = 0.95$$

on sait

$$\Pr\left(-1.96 \leq \frac{\overline{X}-100}{\frac{4}{\sqrt{n}}} \leq 1.96\right) = 0.95$$

On va calculer $Z = \frac{\overline{X}-100}{\frac{4}{\sqrt{n}}}$ avec les données:

```
taille.moyenne ← mean(imcenfant$taille)
n ← length(imcenfant$taille)
Z ← (taille.moyenne − 100)/(4/sqrt(n))
Z
```

```
[1] 2.305572
```

Conclusion

Comme $Z = 2.3055719$ est dehors de l'intervalle $[-1.96, 1.96]$ alors on va rejeter l'hypothèse H_0 avec un niveau de confiance de le 5%. En conséquence on ne peut pas rejeter, avec un niveau de confiance de le 5%, l'hypothèse que la taille moyenne chez la population est de 100 cm.

La question scientifique est maintenant la suivante: En supposant que `taille` $= X \sim \mathcal{N}(\mu, \sigma)$ est-ce la taille moyenne chez la population est différent de 100 cm? On sait

$$\frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} \sim t_{n-1},$$

ou $X_1, \ldots, X_n$ est un échantillon de $X \sim \mathcal{N}(\mu, \sigma)$ et $S^2 = \mathtt{var}(\mathtt{taille})$.

Test d'hypothèse

$$H_0 : \mu = 100 \text{ cm},$$

avec le 95%

$$H_1 : \mu \neq 100 \text{ cm}.$$

avec 5%.

On a:

```
alpha ← 0.05
n ← length(imcenfant$taille)
x.alpha ← qt(1-alpha/2,n-1)
x.alpha
```

```
[1] 1.975799
```

Alors

Si H_0 est vrai, comme

$$\Pr\left(-1.9757989 \leq \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} \leq 1.9757989\right) = 0.95$$

on sait

$$\Pr\left(-1.9757989 \leq \frac{\overline{X} - 100}{\frac{S}{\sqrt{152}}} \leq 1.9757989\right) = 0.95$$

On va calculer $t = \frac{\overline{X}-100}{\frac{S}{\sqrt{152}}}$ avec les données:

```
taille.moyenne ← mean(imcenfant$taille)
taille.sd ← sd(imcenfant$taille)
n ← length(imcenfant$taille)
t ← (taille.moyenne − 100)/(taille.sd/sqrt(n))
t
```

```
[1] 2.177869
```

Conclusion

Comme $t = 2.1778689$ est dehors de l'intervalle $[-1.9757989, 1.9757989]$ alors on va rejeter l'hypothèse H_0 avec un niveau de confiance de le 5%. En conséquence on ne peut pas rejeter, avec un niveau de confiance de le 5%, l'hypothèse que la taille moyenne chez la population est de 100 cm.

Propriété

Si X et Y sont indépendantes alors

$$X - Y \sim \mathcal{N}\left(\mu_X - \mu_Y, \sqrt{\sigma_X^2 + \sigma_Y^2}\right).$$

Soit $X_1, \ldots, X_{n_1} \sim X$ et $Y_1, \ldots, Y_{n_2} \sim Y$ alors

$$\overline{X} - \overline{Y} \sim \mathcal{N}\left(\mu_X - \mu_Y, \sqrt{\frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}}\right)$$

La question scientifique est maintenant la suivante: En supposant que

$$\texttt{taille.F} = X \sim \mathcal{N}(\mu_F, \sigma_F),$$

$$\texttt{taille.G} = Y \sim \mathcal{N}(\mu_G, \sigma_G),$$

avec $\sigma_F = \sigma_G = 4$ cm. Est-ce la taille moyenne chez les filles est différent de la taille moyenne chez les garçons?

Test d'hypothèse

$$H_0 : \mu_F = \mu_G \text{ cm},$$

avec le 95%

$$H_1 : \mu_F \neq \mu_G \text{ cm}.$$

avec 5%.

Maintenant

$$Z = \frac{(\overline{X} - \overline{Y}) - (\mu_F - \mu_G)}{\sqrt{\frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}}} \sim \mathcal{N}(0,1)$$

et sous l'hypothèse H_0 on sait que

$$Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}}} \sim \mathcal{N}(0,1)$$

Si H_0 est vrai, comme

$$\Pr\left(-1.959964 \leq \frac{\overline{X}-\overline{Y}}{\sqrt{\frac{4^2}{n_1}+\frac{4^2}{n_2}}} \leq 1.959964\right) = 0.95$$

On va calculer $Z = \frac{\overline{X}-\overline{Y}}{\sqrt{\frac{4^2}{n_1}+\frac{4^2}{n_2}}}$ avec les données:

```
moyenne.taille.F <- mean(imcenfant$taille[imcenfant$SEXE=="F"])
moyenne.taille.G <- mean(imcenfant$taille[imcenfant$SEXE=="G"])
n.F <- length(imcenfant$taille[imcenfant$SEXE=="F"])
n.G <- length(imcenfant$taille[imcenfant$SEXE=="G"])
c(n.F, n.G)
```

```
[1] 71 81
```

```
Z <- (moyenne.taille.F-moyenne.taille.G)/sqrt(4^2/n.F + 4^2/n.G)
Z
```

```
[1] -2.666609
```

Conclusion

Comme $Z = -2.6666087$ est dehors de l'intervalle $[-1.959964, 1.959964]$ alors on va rejeter l'hypothèse H_0 avec un niveau de confiance de le 5%. En conséquence on ne peut pas rejeter, avec un niveau de confiance de le 5%, l'hypothèse que la taille moyenne chez la population des filles est différent de la taille moyenne chez la population des garçons.

Propriété

Si X et Y sont indépendantes alors

$$X - Y \sim \mathcal{N}\left(\mu_X - \mu_Y, \sqrt{\sigma_X^2 + \sigma_Y^2}\right).$$

Soit $X_1, \ldots, X_{n_1} \sim X$ et $Y_1, \ldots, Y_{n_2} \sim Y$ alors

$$\frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{S_X^2}{n_1} + \frac{S_Y^2}{n_2}}} \sim t_{n_1+n_2-2}$$

La question scientifique est maintenant la suivante: En supposant que

$$\texttt{taille.F} = X \sim \mathcal{N}(\mu_F, \sigma_F),$$

$$\texttt{taille.G} = Y \sim \mathcal{N}(\mu_G, \sigma_G).$$

Est-ce la taille moyenne chez les filles est différent de la taille moyenne chez les garçons?

Test d'hypothèse

$$H_0 : \mu_F = \mu_G \text{ cm},$$

avec le 95%

$$H_1 : \mu_F \neq \mu_G \text{ cm}.$$

avec 5%.

Sous l'hypothèse H_0

$$\frac{\overline{X}-\overline{Y}}{\sqrt{\frac{S_X^2}{n_1}+\frac{S_Y^2}{n_2}}} \sim t_{n_1+n_2-2}$$

et soit

$$x_\alpha = \mathrm{qt}(1-\alpha/2, n_1+n_2-2) \Leftrightarrow \Pr\left(-x_\alpha \leq \frac{\overline{X}-\overline{Y}}{\sqrt{\frac{S_X^2}{n_1}+\frac{S_Y^2}{n_2}}} \leq x_\alpha\right) = 1-\alpha.$$

```
x.alpha ← qt(1−0.05/2,n.F+n.G−2)
x.alpha
```

```
[1] 1.975905
```

On calcule, sous H_0 vrai,

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

```
var.taille.F <- var(imcenfant$taille[imcenfant$SEXE=="F"])
var.taille.G <- var(imcenfant$taille[imcenfant$SEXE=="G"])
t <- (moyenne.taille.F-moyenne.taille.G)/
sqrt(var.taille.F/n.F + var.taille.G/n.G)
t
```

```
[1] -2.574404
```

Conclusion

Comme $t = -2.5744042$ est dehors de l'intervalle $[-1.9759053, 1.9759053]$ alors on va rejeter l'hypothèse H_0 avec un niveau de confiance de le 5%. En conséquence on ne peut pas rejeter, avec un niveau de confiance de le 5%, l'hypothèse que la taille moyenne chez la population des filles est différent de la taille moyenne chez la population des garçons.