Test de Student

Stéphane Canu
stephane.canu@litislab.eu

M8 - Principes du traitement de l'information

June 9, 2017

Plan

L'exemple de l'effet d'un médicament

patient	Groupe	Pression sanguine
t1	traitement	88
t2	traitement	83
t3	traitement	82
t4	traitement	101
t5	traitement	99
t6	traitement	85
t7	traitement	87
t8	traitement	89
t9	traitement	88
p10	placebo	88
p11	placebo	82
p12	placebo	101
p13	placebo	106
p14	placebo	96
p15	placebo	92
p16	placebo	112
p17	placebo	97
	qualitative	quantitative

Question : le traitement fait-il diminuer *significativement* la pression sanguine ?

les hypothèses :

$$\begin{cases} \mathcal{H}_0 : \text{le traitement est inefficace} \\ \mathcal{H}_1 : \text{le traitement la fait baisser} \end{cases}$$

Réponse : comparer les deux échantillons à travers la différence entre leurs moyennes

$$\overline{x}_t - \overline{x}_p = 90,2 - 96,7 = -6,5$$

La question posée se résume ainsi

cette valeur de -6,5 peut elle s'expliquer par un hasard raisonnable ?

Un hasard raisonnable

$\overline{x}_t - \overline{x}_p = -6,5$ peut elle s'expliquer par un hasard raisonnable ?

$$\overline{x}_t = \frac{1}{n_t} \sum_{i=1}^{n_t=9} x_{ti} \qquad \overline{x}_p = \frac{1}{n_p} \sum_{i=1}^{n_p} x_{pi}$$

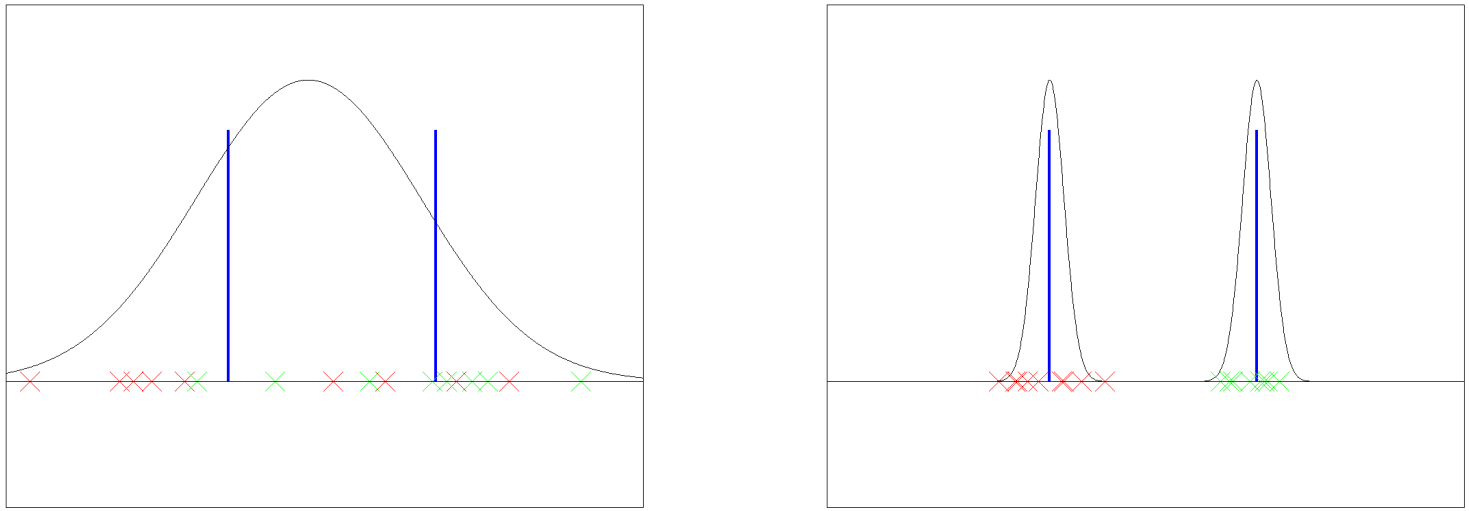


Figure : Illustration des eux cas de figure. Dans le premier cas (à gauche) la variance est grande et donc la distance de 6.5 est petite et due au hasard. Dans le second cas (à droite) la variance est petite et la distance de 6,5 est grande.

pour répondre...

...il faut prendre en compte la variance

Prendre en compte la variance : le modèle

Les trois hypothèses

1. l'hypothèse gaussiènne :
   - mesure des patients avec traitement : $X_t \sim \mathcal{N}(\mu_t, \sigma^2)$
   - mesure des patients sous placébo : $X_p \sim \mathcal{N}(\mu_p, \sigma^2)$
2. même variance : $\sigma_t^2 = \sigma_p^2 = \sigma^2$
3. avec la variance connue donc par exemple : $\sigma^2 = 60$.

les hypothèses :

$$\begin{cases} \mathcal{H}_0 : \text{inefficace} & \mu_t = \mu_p \\ \mathcal{H}_1 : \text{la pression baisse} & \mu_t < \mu_p \end{cases}$$

Nous savons que les moyennes des échantillons suivent une loi normale

- moyenne avec traitement : $\overline{X}_t \sim \mathcal{N}(\mu_t, \frac{\sigma^2}{n_t})$
- moyenne sous placébo : $\overline{X}_p \sim \mathcal{N}(\mu_p, \frac{\sigma^2}{n_p})$

car

$$\begin{aligned} \mathbb{E}(\overline{X}) &= \mathbb{E}(\tfrac{1}{n}\textstyle\sum_{i=1}^n X_i) \\ &= \tfrac{1}{n}\textstyle\sum_{i=1}^n \mathbb{E}(X_i) \\ &= \tfrac{1}{n}\textstyle\sum_{i=1}^n \mu \qquad = \mu \end{aligned}$$

$$\begin{aligned} V(\overline{X}) &= V(\tfrac{1}{n}\textstyle\sum_{i=1}^n X_i) \\ &= \tfrac{1}{n^2}\textstyle\sum_{i=1}^n V(X_i) \\ &= \tfrac{1}{n^2}\textstyle\sum_{i=1}^n \sigma^2 \qquad = \frac{\sigma^2}{n} \end{aligned}$$

Prendre en compte la variance : le modèle

Les trois hypothèses

1. l'hypothèse gaussienne :
   - mesure des patients avec traitement : $X_t \sim \mathcal{N}(\mu_t, \sigma^2)$
   - mesure des patients sous placébo : $X_p \sim \mathcal{N}(\mu_p, \sigma^2)$
2. même variance : $\sigma_t^2 = \sigma_p^2 = \sigma^2$
3. avec la variance connue donc par exemple : $\sigma^2 = 60$.

les hypothèses :

$$\begin{cases} \mathcal{H}_0 : \text{inefficace} & \mu_t = \mu_p \\ \mathcal{H}_1 : \text{la pression baisse} & \mu_t < \mu_p \end{cases}$$

Nous savons que les moyennes des échantillons suivent une loi normale

- moyenne avec traitement : $\overline{X}_t \sim \mathcal{N}(\mu_t, \frac{\sigma^2}{n_t})$
- moyenne sous placébo : $\overline{X}_p \sim \mathcal{N}(\mu_p, \frac{\sigma^2}{n_p})$

La différence des moyennes suit aussi une loi normale :

$$\overline{X}_t - \overline{X}_p \sim \mathcal{N}\left(\mu_t - \mu_p, \sigma^2(\frac{1}{n_t} + \frac{1}{n_p})\right)$$

Le test 1(variance connue)

Le modèle : $\overline{X}_t - \overline{X}_p \sim \mathcal{N}\left(\mu_t - \mu_p, \sigma^2\left(\frac{1}{n_t} + \frac{1}{n_p}\right)\right)$

Le test se rapporte aux deux hypothèses suivantes :

$$\begin{cases} \mathcal{H}_0 : \text{le traitement n'a pas d'effet} & \mu_t - \mu_p = 0 \\ \mathcal{H}_1 : \text{le traitement est efficace} & \mu_t - \mu_p < 0 \end{cases}$$

Maintenant nous faisons l'hypothèse que le traitement n'a pas d'effet.

$$\text{sous } \mathcal{H}_0 : \qquad U = \frac{\overline{X}_t - \overline{X}_p}{\sqrt{\sigma^2\left(\frac{1}{n_t} + \frac{1}{n_p}\right)}} \sim \mathcal{N}(0,1)$$

Avec les données dont nous disposons nous pouvons calculer

$$u = \frac{90,2 - 96,7}{\sqrt{60\left(\frac{1}{9} + \frac{1}{8}\right)}} = -1.73$$

-1,73 est-ce grand ou petit ?

Le test 2 (variance connue)

$$\text{sous } \mathcal{H}_0 : \qquad U = \frac{\overline{X}_t - \overline{X}_p}{\sqrt{\sigma^2\left(\frac{1}{n_t} + \frac{1}{n_p}\right)}} \sim \mathcal{N}(0,1)$$

Avec les données dont nous disposons nous pouvons calculer

$$u = \frac{90,2 - 96,7}{\sqrt{60\left(\frac{1}{9} + \frac{1}{8}\right)}} = -1.73$$

En prenant les tables de la loi normale nous constatons que

$$\mathbb{P}(U \leq -1.7343) = 0,041$$

Il y a donc moins de 5% de chances d'observer un tel résultat. Il ne nous apparait donc pas raisonnable d'expliquer cette différence entre le moyennes par le hasard seul. Nous concluons dans ce cas en rejetant cette hypothèse. Il nous semble plus raisonnable d'admettre que le traitement à un effet.

Récapitulons : le test de comparaison des moyennes

1. la question : les deux groupes sont ils des réalisation de la même loi
2. le modèle : gaussien
3. les hypothèses : même variance σ^2 connue
4. caclul de

$$u = \frac{\overline{x}_t - \overline{x}_p}{\sqrt{\sigma^2\left(\frac{1}{n_t} + \frac{1}{n_p}\right)}}$$

$\overline{x}_t$ moyenne avec traitement
$\overline{x}_p$ moyenne sans traitement
n_t nombre de cas avec traitement
n_p nombre de cas sans traitement

5. calcul de la p-valeur $U \sim \mathcal{N}(0,1)$ (ou lecture sur les tables)

$$pval = \mathbb{P}(U \leq u)$$

6. on décide qu'on ne peut pas conclure à l'efficacité du traitement si la p-valeur est supérieure à 0,05, si $pval \geq 0,05$

Les trois variantes :

la pression :

diminue augmente varie .

$$\begin{cases} \mathcal{H}_0 : \mu_t - \mu_p = 0 \\ \mathcal{H}_1 : \mu_t - \mu_p < 0 \end{cases} \qquad \begin{cases} \mathcal{H}_0 : \mu_t - \mu_p = 0 \\ \mathcal{H}_1 : \mu_t - \mu_p > 0 \end{cases} \qquad \begin{cases} \mathcal{H}_0 : \mu_t - \mu_p = 0 \\ \mathcal{H}_1 : \mu_t - \mu_p \neq 0 \end{cases}$$

pval =

$\mathbb{P}(U \leq u)$ $\mathbb{P}(U \geq u)$ $\mathbb{P}(U \leq -|u|) + \mathbb{P}(U \geq |u|)$

quand la question change...

le calcul de la *pval* change

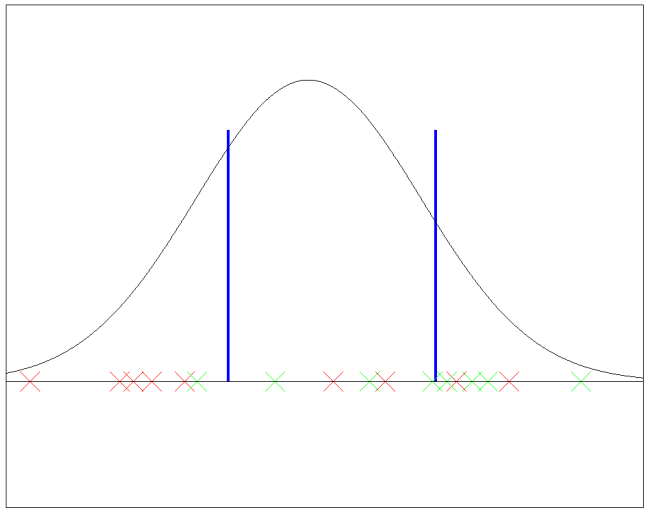
Exemple : pour $u = -1,73$, *pval* =

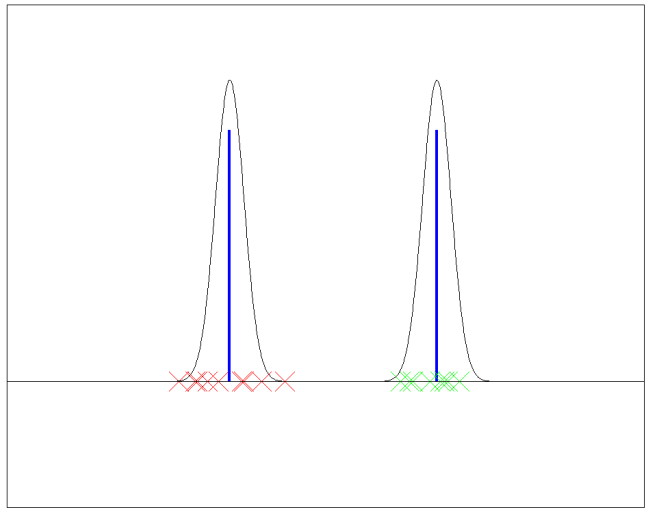
dim : $\mathbb{P}(U \leq -1,73) = 0,041$

aug : $\mathbb{P}(U \geq -1,73) = 1 - 0,041 = 0,959$

var : $\mathbb{P}(U \leq -1,73) + \mathbb{P}(U \geq 1,73) = 0,041 + 0,041 = 0,082$

une interprétation de la statistique





$$u = \frac{\text{signal}}{\text{bruit}}$$

$$= \frac{\text{écart entre les moyennes des deux groupes}}{\text{variabilité des observations}}$$

$$= \frac{\overline{x}_t - \overline{x}_p}{\sqrt{\sigma^2\left(\frac{1}{n_t} + \frac{1}{n_p}\right)}}$$

Plan

Si la variance est inconnue

Dans ce cas on remplace la variance inconnue σ^2 par son estimateur $\widehat{\sigma}^2$. En conséquence la nouvelle variable aléatoire ainsi construite n'est plus distribué selon une loi normale mais suit une loi et Student à $n_t + n_p - 2$ degrés de liberté.

$$T_{n_t+n_p-2} = \frac{\overline{X}_t - \overline{X}_p}{\sqrt{\widehat{\sigma}^2\left(\frac{1}{n_t} + \frac{1}{n_p}\right)}} \sim \mathcal{T}_{n_t+n_p-2}$$

avec $\widehat{\sigma}^2 = \frac{1}{n_t+n_p-2}\left(\sum_{i=1}^{n_t}(X_{ti} - \overline{X}_t)^2 + \sum_{i=1}^{n_p}(X_{pi} - \overline{X}_p)^2\right)$.

$$t = \frac{90,2 - 96,7}{\sqrt{63,4\left(\frac{1}{9} + \frac{1}{8}\right)}} = -1.68$$

En prenant les tables de la loi de Student nous constatons que

$$pval = \mathbb{P}(T_{n_t+n_p-2} \leq -1.68) = 0,056$$

Il y a dans ce cas plus de 5% de chances d'observer un tel résultat. Il nous apparait donc plausible d'expliquer cette différence entre le moyennes par le seul effet du hasard. Nous concluons dans ce cas en gardant cette hypothèse. Il n'y a pas assez d'évidence expérimentale pour nous convaincre que le traitement a vraiment un effet. Si le médecin souhaite poursuivre, il lui faut refaire une expérience sur plus de sujets.

Récapitulons : le test de comparaison des moyennes

1. la question : les deux groupes sont ils des réalisation de la même loi
2. le modèle : gaussien
3. les hypothèses : même variance σ^2 inconnue
4. caclul de

$$t = \frac{\overline{x}_t - \overline{x}_p}{\sqrt{\widehat{\sigma}^2\left(\frac{1}{n_t} + \frac{1}{n_p}\right)}}$$

$$\widehat{\sigma}^2 = \frac{1}{n_t+n_p-2}\left(\sum_{i=1}^{n_t}(x_{ti} - \overline{x}_t)^2 + \sum_{i=1}^{n_p}(x_{pi} - \overline{x}_p)^2\right)$$

$\overline{x}_t$ moyenne avec traitement
$\overline{x}_p$ moyenne sans traitement
n_t nombre de cas avec traitement
n_p nombre de cas sans traitement

5. calcul de la p-valeur $T \sim \mathcal{T}_{n_t+n_p-2}$ (ou lecture sur les tables)

$$pval = \mathbb{P}(T \leq t)$$

6. on décide qu'on ne peut pas conclure à l'efficacité du traitement si la p-valeur est supérieure à 0,05, si $pval \geq 0,05$

Plan

La loi de Student : définition

- Soit $N \sim \mathcal{N}(0,1)$ une variable aléatoire normale centrée réduite.

- Soit X_n la variable aléatoire distribuée suivant une loi du χ^2 à n ddl
  - C'est le cas par exemple, si $N_1, N_2, ..., N_n$ un échantillon de n réalisation i.i.d. une variable aléatoire normale centrée réduite quand $X_n = \sum_{i=1}^{n} N_i^2$

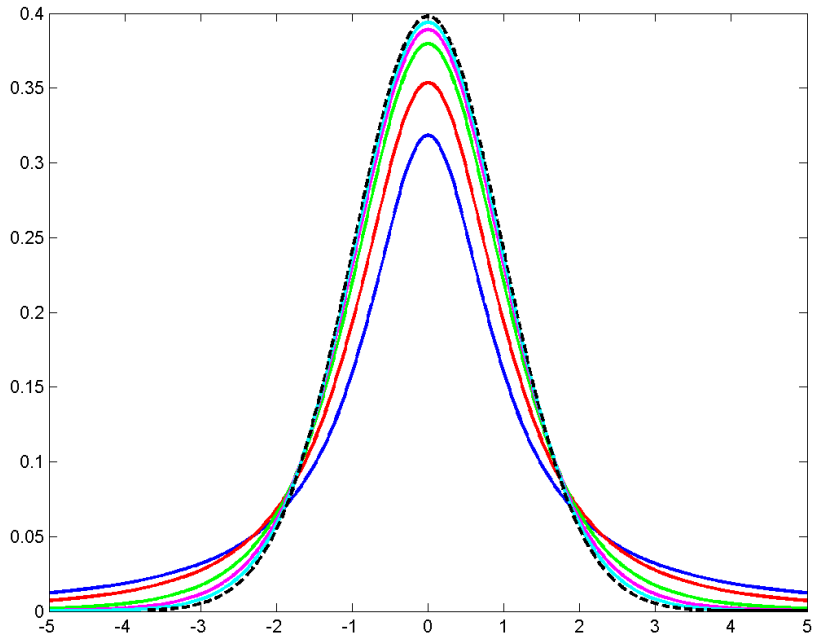
- supposons que N et X_n sont indépendantes $(i.e.\ cov(Y, X_n) = 0)$

Definition (La loi de student)

On appelle loi de student à n degrés de libertés la loi de la variable aléatoire T_n

$$T_n = \frac{N}{\sqrt{\frac{X_n}{n}}} \qquad N \sim \mathcal{N}(0,1) \qquad X_n \sim \chi^2_n$$

La loi de Student : $T_n = \frac{N}{\sqrt{\frac{X_n}{n}}}$



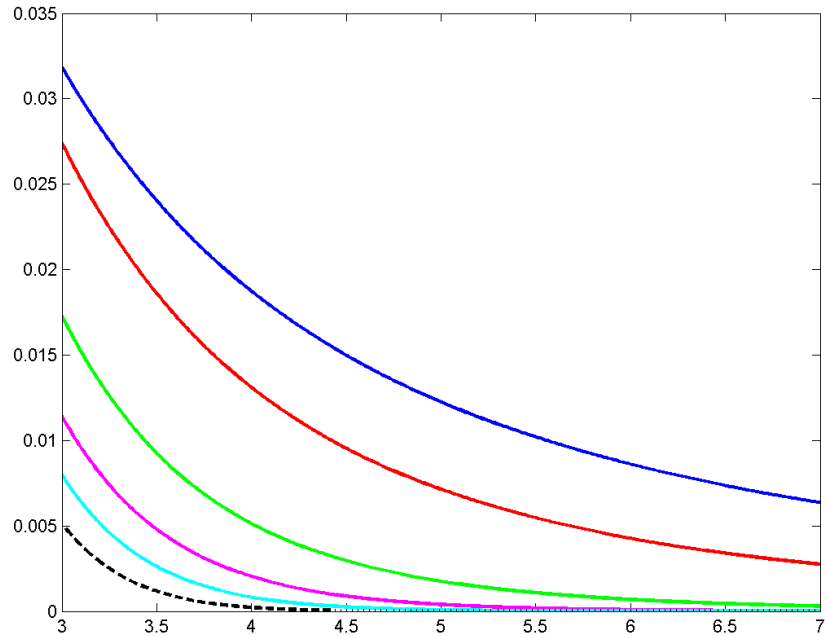


Figure : Exemples de loi de student pour 1 (bleu), 2 (rouge), 5 (vert), 10 (violet) et 20 (bleu ciel) degrés de liberté. La courbe en pointillés noir est la courbe de Gauss donnée comme référence. La figure de droite montre un zoom sur la « queue » de la distribution.

Loi de Student et loi normale

$$T_n \xrightarrow[n \to +\infty]{} \mathcal{N}(0,1)$$

Propriétés et approximation

- Publiée pour la première fois en 1908 par William Sealy Gosset qui travaillait chez Guinness (la brasserie de Dublin). Pour des raisons commerciales, il a du utiliser le pseudonyme de Student, qui restera attaché à cette loi.

- tend vers une loi normale n > 30

- attention la différence est plus importante dans les « queue » de la distribution :
  - $N \sim \mathcal{N}(0,1)$: $\mathbb{P}(N > 2) = 0,023$ `p1 = 1-cdf('norm',2,0,1)`
  - $T \sim \mathcal{T}_1$: $\mathbb{P}(T > 2) = 0,148$ `p2 = 1-cdf('t',2,1)`
  - $T \sim \mathcal{T}_2$: $\mathbb{P}(T > 2) = 0,092$ `p2 = 1-cdf('t',2,2)`
  - $T \sim \mathcal{T}_{10}$: $\mathbb{P}(T > 2) = 0,038$ `p2 = 1-cdf('t',2,10)`

$$U \sim \mathcal{N}(0,\sigma^2) \qquad N = \frac{U}{\sigma} \sim \mathcal{N}(0,1) \qquad T = \frac{N}{\widehat{\sigma}} = \frac{N}{\sqrt{\frac{N_1^2+N_2^2}{2}}} \sim \mathcal{T}_2$$

Le cas de la moyenne d'un échantillon gaussien

Soit $X \sim \mathcal{N}(\mu, \sigma^2)$ une variable aléatoire normale d'espérance μ et de variance σ^2. Soit $X_1, X_2, ..., X_n$ un échantillon de n réalisation i.i.d. de cette variable aléatoire. La moyenne $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ de cet échantillon suit aussi une loi normale

$$\overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

car $\mathbb{E}(\overline{X}) = \mu$ et $V(\overline{X}) = \frac{\sigma^2}{n}$:

$$\begin{aligned} \mathbb{E}(\overline{X}) &= \mathbb{E}(\tfrac{1}{n}\textstyle\sum_{i=1}^{n} X_i) \\ &= \tfrac{1}{n}\textstyle\sum_{i=1}^{n} \mathbb{E}(X_i) \\ &= \tfrac{1}{n}\textstyle\sum_{i=1}^{n} \mu \\ &= \mu \end{aligned} \qquad \begin{aligned} V(\overline{X}) &= V(\tfrac{1}{n}\textstyle\sum_{i=1}^{n} X_i) \\ &= \tfrac{1}{n^2}\textstyle\sum_{i=1}^{n} V(X_i) \\ &= \tfrac{1}{n^2}\textstyle\sum_{i=1}^{n} \sigma^2 \\ &= \frac{\sigma^2}{n} \end{aligned}$$

Le cas de la moyenne d'un échantillon gaussien

Soit $X \sim \mathcal{N}(\mu, \sigma^2)$ une variable aléatoire normale d'espérance μ et de variance σ^2. Soit $X_1, X_2, ..., X_n$ un échantillon de n réalisation i.i.d. de cette variable aléatoire. La moyenne $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ de cet échantillon suit aussi une loi normale

$$\overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

On peut donc construire la variable normale centrée réduite

$Y = \frac{\overline{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}} \sim \mathcal{N}(0,1)$. Or $Z_{n-1} = \sum_{i=1}^{n} \frac{(X_i-\overline{X})^2}{\sigma^2} \sim \chi^2_{n-1}$

On peut construire une variable aléatoire suivant une loi de Student

$$T_{n-1} = \frac{Y}{\sqrt{\frac{Z_{n-1}}{n-1}}} = \frac{\frac{\overline{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}}}{\sqrt{\frac{\sum_{i=1}^{n}\frac{(X_i-\overline{X})^2}{\sigma^2}}{n-1}}} = \frac{\overline{X}-\mu}{\sqrt{\frac{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2}{n}}} = \frac{\overline{X}-\mu}{\frac{S_{n-1}}{\sqrt{n}}}$$

avec $S^2_{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$.

Le test de Student (t-test) : deux échantillons gaussien

Soit $X \sim \mathcal{N}(\mu_x, \sigma^2)$ et $Y \sim \mathcal{N}(\mu_y, \sigma^2)$ deux loi de même variance.

On tire deux échantillons suivant ces deux loi.
Soient $X_1, ..., X_{n_x}$ et $Y_1, ..., Y_{n_y}$ ces deux échantillons.
Les variables suivantes $\overline{X} = \frac{1}{n}\sum_{i=1}^{n_x} X_i$ et $S_x^2 = \sum_{i=1}^{n_x}(X_i - \overline{X})^2$ sont caractérisées par les lois :

$$\overline{X} \sim \mathcal{N}\left(\mu_x, \frac{\sigma^2}{n_x}\right) ; \quad \overline{Y} \sim \mathcal{N}\left(\mu_y, \frac{\sigma^2}{n_y}\right) ; \quad \frac{S_x^2}{\sigma^2} \sim \chi^2_{n_x-1} ; \quad \frac{S_y^2}{\sigma^2} \sim \chi^2_{n_y-1}$$

et donc

$$\overline{X} - \overline{Y} \sim \mathcal{N}\left(\mu_x - \mu_y, (\frac{1}{n_x} + \frac{1}{n_y})\sigma^2\right) ; \qquad \frac{S_x^2}{\sigma^2} + \frac{S_y^2}{\sigma^2} \sim \chi^2_{n_x+n_y-2}$$

Le test de Student (t-test)

$$\overline{X} - \overline{Y} \sim \mathcal{N}\left(\mu_x - \mu_y, (\frac{1}{n_x} + \frac{1}{n_y})\sigma^2\right) ; \qquad \frac{S_x^2}{\sigma^2} + \frac{S_y^2}{\sigma^2} \sim \chi^2_{n_x+n_y-2}$$

On définit alors la variable de Student suivante :

$$T_{n_x+n_y-2} = \sqrt{n_x + n_y - 2}\, \frac{\overline{X} - \overline{Y} - (\mu_x - \mu_y)}{\sqrt{\left(\frac{1}{n_x} + \frac{1}{n_y}\right) S_{xy}^2}}$$

avec $S_{xy}^2 = S_x^2 + S_y^2 = \sum_{i=1}^{n_x}(X_i - \overline{X})^2 + \sum_{i=1}^{n_y}(Y_i - \overline{Y})^2$

Si l'on fait l'hypothèse que $\mu_x = \mu_y$

$$T = \sqrt{n_x + n_y - 2}\, \frac{\overline{X} - \overline{Y}}{\sqrt{\left(\frac{1}{n_x} + \frac{1}{n_y}\right) S_{xy}^2}}$$

suit une loi de Student à $n_x + n_y - 2$ degrés de liberté.

Plan

Le test de Student (t-test)

les deux échantilons : $X_{t1}, ..., X_{ti}, ..., X_{tn_t}, X_{p1}, ..., X_{pi}, ..., X_{pn_p}$ i.i.d

Les deux hypothèses

1. l'hypothèse gaussienne :
   - soit $X_{ti} \sim \mathcal{N}(\mu_t, \sigma^2)$
   - et $X_{pi} \sim \mathcal{N}(\mu_p, \sigma^2)$
2. même variance : $\sigma_t^2 = \sigma_p^2 = \sigma^2$

la question : les deux échantillons que nous observons sont-ils des réalisations d'une même variable aléatoire ?

les hypothèses :
$$\begin{cases} \mathcal{H}_0 : \text{échantillons de même loi} & \mu_t = \mu_p \\ \mathcal{H}_1 : \text{de lois différentes} & \mu_t > \mu_p \end{cases}$$

la statistique :
$$T = \frac{\overline{X}_t - \overline{X}_p}{\sqrt{\widehat{\sigma}^2\left(\frac{1}{n_t} + \frac{1}{n_p}\right)}} \sim \mathcal{T}_{n_t+n_p-2}$$

avec $\widehat{\sigma}^2 = \frac{1}{n_t+n_p-2}\left(\sum_{i=1}^{n_t}(X_{ti} - \overline{X}_t)^2 + \sum_{i=1}^{n_p}(X_{pi} - \overline{X}_p)^2\right)$

Mise en œuvre du test de student

1. caclul de
$$\overline{x}_t = \frac{1}{n_t}\sum_{i=1}^{n_t} x_{ti} \quad \text{moyenne avec traitement}$$
$$\overline{x}_p = \frac{1}{n_p}\sum_{i=1}^{n_p} x_{pi} \quad \text{moyenne sans traitement}$$
2. caclul de $\widehat{\sigma}^2 = \frac{1}{n_t+n_p-2}\left(\sum_{i=1}^{n_t}(x_{ti}-\overline{x}_t)^2 + \sum_{i=1}^{n_p}(x_{pi}-\overline{x}_p)^2\right)$
3. caclul de
$$t = \frac{\overline{x}_t - \overline{x}_p}{\sqrt{\widehat{\sigma}^2\left(\frac{1}{n_t}+\frac{1}{n_p}\right)}}$$
n_t nombre de cas avec traitement
n_p nombre de cas sans traitement
4. calcul du nombre de degrés de liberté $d = n_t + n_p - 2$
5. calcul de la p-valeur $T \sim \mathcal{T}_d$ (ou lecture sur les tables)
$$pval = \mathbb{P}(T \geqslant t)$$
6. on décide qu'on ne peut pas conclure à l'efficacité du traitement si la p-valeur est supérieure à 0,05, si $pval \geqslant 0,05$

Exemple de mise en œuvre du test de student

groupe avec traitement (t)	30.02	29.99	30.11	29.97	30.01	29.99
groupe sans traitement (p)	29.89	29.93	29.72	29.98	30.02	29.98

Question : le traitement augmente-t-il la mesure ?

Exemple de mise en œuvre du test de student

groupe avec traitement (t)	30.02	29.99	30.11	29.97	30.01	29.99
groupe sans traitement (p)	29.89	29.93	29.72	29.98	30.02	29.98

Question : le traitement augmente-t-il la mesure ?

Réponse : on effectue le test de student :

1. $\overline{x}_t = 30.015, \quad \overline{x}_p = 29.92 \qquad \overline{x}_t - \overline{x}_p = 0.095$
2. $\widehat{\sigma}^2 = \frac{1}{10}\left(\sum_{i=1}^{6}(x_{ti} - 30.015)^2 + \sum_{i=1}^{6}(x_{pi} - 29.92)^2\right) \approx 0.0071$
3. $$t = \frac{\overline{x}_t - \overline{x}_p}{\sqrt{\widehat{\sigma}^2\left(\frac{1}{n_t} + \frac{1}{n_p}\right)}} \approx \frac{0.095}{\sqrt{0.0071\left(\frac{1}{6} + \frac{1}{6}\right)}} = 1.959$$
4. calcul du nombre de degrés de liberté $d = n_t + n_p - 2 = 10$
5. calcul de la p-valeur $T \sim \mathcal{T}_d$ (ou lecture sur les tables)

$$pval = \mathbb{P}(T \geq 1.959) = \texttt{1-cdf('t',1.959,10)} = 0.0393$$

6. on décide qu'on peut conclure à l'efficacité du traitement car la p-valeur est inférieure à 0,05.

```
>> help ttest2

 TTEST2 Two-sample T-test with pooled or unpooled variance estimate.
    H = TTEST2(X,Y) performs a T-test of the hypothesis that two
    independent samples, in the vectors X and Y, come from distributions
    with equal means, and returns the result of the test in H.  H=0
    indicates that the null hypothesis ("means are equal") cannot be
    rejected at the 5% significance level.  H=1 indicates that the null
    hypothesis can be rejected at the 5% level.  The data are assumed to
    come from normal distributions with unknown, but equal, variances.  X
    and Y can have different lengths.

    X and Y can also be matrices or N-D arrays.  For matrices, TTEST2
    performs separate T-tests along each column, and returns a vector of
    results.  X and Y must have the same number of columns.  For N-D
    arrays, TTEST2 works along the first non-singleton dimension.  X and Y
    must have the same size along all the remaining dimensions.

    TTEST2 treats NaNs as missing values, and ignores them.

    H = TTEST2(X,Y,ALPHA) performs the test at the significance level
    (100*ALPHA)%.  ALPHA must be a scalar.

    H = TTEST2(X,Y,ALPHA,TAIL) performs the test against the alternative
    hypothesis specified by TAIL:
        'both'  -- "means are not equal" (two-tailed test)
        'right' -- "mean of X is greater than mean of Y" (right-tailed test)
        'left'  -- "mean of X is less than mean of Y" (left-tailed test)
    TAIL must be a single string.

    H = TTEST2(X,Y,ALPHA,TAIL,VARTYPE) allows you to specify the type of
    test.  When VARTYPE is 'equal', TTEST2 performs the default test
    assuming equal variances.  When VARTYPE is 'unequal', TTEST2 performs
    the test assuming that the two samples come from normal distributions
    with unknown and unequal variances.  This is known as the Behrens-Fisher
    problem. TTEST2 uses Satterthwaite's approximation for the effective
    degrees of freedom.  VARTYPE must be a single string.

    [H,P] = TTEST2(...) returns the p-value, i.e., the probability of
    observing the given result, or one more extreme, by chance if the null
    hypothesis is true.  Small values of P cast doubt on the validity of
    the null hypothesis.

    [H,P,CI] = TTEST2(...) returns a 100*(1-ALPHA)% confidence interval for
    the true difference of population means.

    [H,P,CI,STATS] = TTEST2(...) returns a structure with f
       'tstat' -- the value of the test statistic
       'df'    -- the degrees of freedom of the test
       'sd'    -- the pooled estimate of the population sta
                  (for the equal variance case) or a vecto
                  unpooled estimates of the population sta
                  (for the unequal variance case)

    [...] = TTEST2(X,Y,ALPHA,TAIL,VARTYPE,DIM) works along
    X and Y.  Pass in [] to use default values for ALPHA, '

    See also ttest, ranksum, vartest2, ansaribradley.
```

Plan

L'exemple de la relation entre oxygène dissout et pression

patient	O_2	Pression sanguine
p1	0,31	88
p2	0,30	83
p3	0,29	82
p4	0,35	101
p5	0,33	99
p6	0,31	85
p7	0,30	87
p8	0,34	89
p9	0,32	88
p10	0,28	88
p11	0,30	82
p12	0,33	101
p13	0,31	106
p14	0,32	96
p15	0,30	92
p16	0,35	112
p17	0,31	97
	quantitative	quantitative

Question : Il y a t'il une relation entre ces deux variables ?

les hypothèses $\begin{cases} \mathcal{H}_0 : \text{indépendance} \\ \mathcal{H}_1 : \text{dépendance} \end{cases}$

Réponse : tester la pente de la droite

$$pression = aO_2 + b + \varepsilon$$

les hypothèses $\begin{cases} \mathcal{H}_0 : a = 0 \\ \mathcal{H}_1 : a \neq 0 \end{cases}$

la regression donne $\widehat{a} = 0,12$

Cette valeur peut elle s'expliquer par un hasard raisonnable ?

un hasard raisonnable...

1. supposons qu'il y a indépendance $a = 0$
2. générons plein ($m = 1000, 1000000, +\infty$) d'échantillons

$$(x_i, y_{ij} = ax_i + b + \varepsilon_{ij}), \quad i = 1, n \;\; j = 1, m$$

3. pour chacun de ces échantillon calculons $\widehat{a}_j$
4. regardons la probabilité $\mathbb{P}(|\widehat{a}| > 0,12)$
5. si cette probabilité est trop petite, il n'est pas « raisonnable » de considérer que l'hypothèse d'indépendance est exacte.

Comparaisons de deux variables quantitatives et régression

$$y_i = ax_i + b + \varepsilon_i$$

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

indépendance des ε_i

$$\widehat{a} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

$$\widehat{a} \sim \mathcal{N}\left(a, \sigma^2 \frac{1}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right)$$

$$\frac{\widehat{a} - a}{\sqrt{\frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}} \sim \mathcal{N}(0, 1)$$

$$\widehat{\varepsilon}_i = y_i - (\widehat{a}x_i + \widehat{b})$$

$$\frac{\varepsilon_i}{\sigma} \sim \mathcal{N}(0, 1) \quad \Rightarrow \quad \frac{1}{\sigma^2}\sum_{i=1}^{n}\varepsilon_i^2 \sim \chi^2_n$$

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}\widehat{\varepsilon}_i^2 \sim \chi^2_{n-2}$$

Pente de la droite de régression et loi de student

$$\frac{\widehat{a}-a}{\sqrt{\frac{\sigma^2}{\sum_{i=1}^n (x_i-\overline{x})^2}}} \sim \mathcal{N}(0,1) \qquad \frac{1}{\sigma^2}\sum_{i=1}^n \widehat{\varepsilon}_i^2 \sim \chi^2_{n-2}$$

or $\frac{\mathcal{N}}{\sqrt{\frac{\chi^2_n}{n}}} \sim \mathcal{T}_n^2$ suit une loi de student à n degrés de libertés

$$\frac{\frac{\widehat{a}-a}{\sqrt{\frac{\sigma^2}{\sum_{i=1}^n (x_i-\overline{x})^2}}}}{\sqrt{\frac{1}{\sigma^2(n-2)}\sum_{i=1}^n \widehat{\varepsilon}_i^2}} \sim \mathcal{T}_{n-2} \qquad \Longrightarrow \qquad \frac{\widehat{a}-a}{\sqrt{\frac{\widehat{\sigma}^2}{S_x^2}}} \sim \mathcal{T}_{n-2}$$

avec $\widehat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^n \left(y_i - (\widehat{a}x_i + \widehat{b})\right)^2$ et $S_x^2 = \sum_{i=1}^n (x_i - \overline{x})^2$

Mise en œuvre du test sur la pente de la régression

1. les hypothèses : $\begin{cases} \mathcal{H}_0 : \text{indépendance} & a = 0 \\ \mathcal{H}_1 : \text{dépendance} & a \neq 0 \end{cases}$
2. caclul de $\widehat{a} = \dfrac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$
3. calcul de $\widehat{\sigma}^2 = \frac{1}{n-2}\displaystyle\sum_{i=1}^{n}\left(y_i - (\widehat{a}x_i + \widehat{b})\right)^2$ et de $S_x^2 = \displaystyle\sum_{i=1}^{n}(x_i - \overline{x})^2$
4. caclul de
$$t = \frac{\widehat{a}}{\sqrt{\frac{\widehat{\sigma}^2}{S_x^2}}}$$
5. calcul du nombre de degrés de liberté $d = n - 2$
6. calcul de la p-valeur $T \sim \mathcal{T}_d$ (ou lecture sur les tables)
$$pval = \mathbb{P}(|T| \geq t)$$
7. on décide qu'on ne peut pas conclure à l'efficacité du traitement si la p-valeur est supérieure à 0,05, si $pval \geq 0,05$

Plan

Conclusion

- La question
  - cette variable quantitative est elle indépendantes de cette variable qualitative ?
  - comparaison de deux échantillons quantitatifs

- il vérifier les hypothèses avant d'effectuer un test de student
  - distribution normale (par exemple un test du χ^2 adapté)
  - égalité de variances (test de Fisher)

  sinon il faut faire un autre test comme celui de Wilcoxon ou de Mann et Whitney

- il existes plusieurs variations du test de student...
  - un échantillon (test d'une valeur de l'espérance) puisque $\frac{\overline{X}-\mu}{\frac{S_{n-1}}{\sqrt{n}}} \sim \mathcal{T}_{n-1}$
  - deux échantillons appariés
  - test de la pente de la régression simple

- Il existe une théorie et des théorèmes pour définir les test
  - théorème de Neyman Pearson

Repéres bibliographiques

- http://en.wikipedia.org/wiki/Student's_t-test
- http://www.iumsp.ch/Enseignement/pregradue/Student.pdf
- http://www.socialresearchmethods.net/kb/stat_t.php
- http://nte-serveur.univ-lyon1.fr/immediato/Math/Enseignement/07%20Statistiques/19.%20Comparaison%20de%20deux%20moyennes%20-%20test%20de%20Student/chapitre_19.htm