Statistiques (TSTL)

1/ Test d'hypothèse
2/ Estimation
3/ Différence significative
4/ Conditions de validité
5/ Un exemple de test d'hypothèse
6/ Un exemple de différence significative

On s'intéresse à la proportion de français favorables au président. Cette proportion est inconnue et on l'appelle *p*.

On peut se poser deux questions :
- on pense connaître la valeur de *p* et on veut tester cette hypothèse ;
- on veut une estimation de *p*.

Dans les deux cas on fait un sondage sur *n* personnes.
Ce sondage donne une proportion (ou fréquence) de sondés favorables au président que l'on appelle *f*.

1/ Test d'hypothèse

On pense connaître *p* et on veut tester cette hypothèse.
L'intervalle de fluctuation des fréquences au seuil de 95 % est

$$I = \left[p - 1{,}96\sqrt{\frac{p(1-p)}{n}} \ ; \ p + 1{,}96\sqrt{\frac{p(1-p)}{n}} \right].$$

Le sondage donne une fréquence *f*.
Si $f \in I$, on estime que notre hypothèse est vraie.
Si $f \notin I$, on rejette notre hypothèse.

Un autre sondage peut donner une autre fréquence *f*. C'est la fluctuation d'échantillonnage.
Dire que l'intervalle de fluctuation est au seuil de 95 % signifie que, si la valeur de *p* est bien exacte, 95 % des sondages donnent une fréquence *f* appartenant à l'intervalle de fluctuation.
Cela signifie aussi que, si la valeur de *p* est bien exacte, *f* a une probabilité de 0,95 d'appartenir à l'intervalle de fluctuation.
Cela signifie aussi que, lorsqu'on fait un sondage, on a une probabilité de 0,05 d'obtenir une fréquence *f* en dehors de l'intervalle de fluctuation.

En revanche, *p* ne dépend pas du sondage donc l'intervalle de fluctuation ne dépend pas du sondage.

François Tuaillon. Mars 2020

2/ Estimation

Un sondage donne une fréquence *f*.
L'intervalle de confiance des fréquences au seuil de 95 % est
$\left[f - 1{,}96\sqrt{\frac{f(1-f)}{n}} \ ; \ f + 1{,}96\sqrt{\frac{f(1-f)}{n}} \right]$.
Cela donne une estimation de la proportion inconnue *p*.

Un autre sondage peut donner une autre fréquence *f*.
Dire que l'intervalle de confiance est au seuil de 95 % signifie que *p* appartient à l'intervalle de confiance avec une probabilité de 95 %.
On dit alors que *p* appartient à l'intervalle de confiance au niveau de confiance de 95 %.

f dépend du sondage donc l'intervalle de confiance dépend du sondage.

3/ Différence significative

Un intervalle de confiance permet aussi de dire si une différence est significative.
On fait un sondage sur une population. On trouve une proportion f_1 et un intervalle de confiance I_1.
On fait un autre sondage sur une autre population. On trouve une proportion f_2 et un intervalle de confiance I_2.
Si f_1 et f_2 sont différents, on peut se demander si cette différence est significative (si les populations sont effectivement différentes).
Si I_1 et I_2 sont disjoints, on dit que la différence est significative.
Si I_1 et I_2 se chevauchent, on dit que la différence n'est pas significative.

C'est rare, mais il est possible que les deux sondages donnent le même résultat : $f_1 = f_2$. Dans ce cas, on dit qu'il n'y a pas de différence significative de proportion dans les deux populations.

4/ Conditions de validité

Au seuil de 95 %, on utilise la constante 1,96. Pour un autre seuil, on utilise une autre constante.
Pour que tout cela fonctionne, l'effectif *n* du sondage doit être suffisamment grand et la probabilité *p* d'un succès ne doit être ni trop proche de 0, ni trop proche de 1.
Les conditions demandées par les sujets d'annales sont $n \geq 30$; $np \geq 5$ et $n(1-p) \geq 5$.

François Tuaillon. Mars 2020

5/ Un exemple de test d'hypothèse

M. le maire d'une ville de 200 000 habitants affirme que 55 % de ses administrés majeurs pratiquent un sport.

I/ Lors d'un sondage, sur 100 habitants majeurs interrogés seulement la moitié affirme pratiquer un sport.
Que penser de l'affirmation de M. le maire ?

II/ On réalise un autre sondage. Sur 1 000 habitants majeurs interrogés seulement la moitié affirme pratiquer un sport.
Que penser de l'affirmation de M. le maire ?

III/ Comparer ces deux réponses.

Soit X la variable aléatoire égale au nombre de sondés déclarant pratiquer un sport.

I/ Avec le sondage sur 100 personnes

1/ Si le sondage est bien réalisés, X suit une loi binomiale.
Les paramètres de cette loi binomiale sont
$n = 100$ car on interroge 100 personne
et $p = 0{,}55$ car la probabilité d'un succès est 0,55.
La probabilité d'un échec est $q = 1 - p = 0{,}45$.
$n = 100 \geq 30$; $n\,p = 55 \geq 5$ et $n\,(1 - p) = 45 \geq 5$ donc on peut chercher un intervalle de fluctuation.

2/ Au seuil de 95 %, l'intervalle de fluctuation de la proportion de sportif est

$$I = \left[\, p - 1{,}96\sqrt{\frac{p(1-p)}{n}} \;;\; p + 1{,}96\sqrt{\frac{p(1-p)}{n}} \,\right]$$

$$= \left[\, 0{,}55 - 1{,}96\sqrt{\frac{0{,}55\times 0{,}45}{100}} \;;\; 0{,}55 + 1{,}96\sqrt{\frac{0{,}55\times 0{,}45}{100}} \,\right]$$

$$= [\, 0{,}45 \;;\; 0{,}65 \,].$$

3/ Que penser des propos de M. le maire ?
M. le maire affirme que la proportion de ses administrés majeurs qui pratiquent un sport est $p = 0{,}55$.
L'intervalle de fluctuation au seuil de 95 % est [0,45 ; 0,65].
La proportion (ou fréquence) de sportifs donnée par le sondage est $f = 0{,}50$.
$0{,}50 \in [\, 0{,}45 \;;\; 0{,}65 \,]$ donc
Au seuil de 95 %, on accepte l'affirmation de M. le maire.

François Tuaillon. Mars 2020

II/ Avec le sondage sur 1 000 personnes

1/ $n = 1\,000 \geq 30$; $n\,p = 550 \geq 5$ et $n\,(1 - p) = 450 \geq 5$ donc on peut chercher un intervalle de fluctuation.

2/ Au seuil de 95 %, l'intervalle de fluctuation de la proportion de sportif est

$$I = \left[\, p - 1{,}96\sqrt{\frac{p(1-p)}{n}} \;;\; p + 1{,}96\sqrt{\frac{p(1-p)}{n}} \,\right]$$

$$= \left[\, 0{,}55 - 1{,}96\sqrt{\frac{0{,}55 \times 0{,}45}{1000}} \;;\; 0{,}55 + 1{,}96\sqrt{\frac{0{,}55 \times 0{,}45}{1000}} \,\right]$$

$$= [\, 0{,}52 \;;\; 0{,}58 \,].$$

3/ Que penser des propos de M. le maire ?

M. le maire affirme que la proportion de ses administrés majeurs qui pratiquent un sport est $p = 0{,}55$.

L'intervalle de fluctuation au seuil de 95 % est [0,52 ; 0,58].

La proportion (ou fréquence) de sportifs donnée par le sondage est $f = 0{,}50$.

$0{,}50 \notin [\, 0{,}52 \;;\; 0{,}58 \,]$ donc

Au seuil de 95 %, on rejette l'affirmation de M. le maire.

Comment cela se fait-il ?

Il y a deux explications

- le sondeur n'a pas eu de chance. Il y avait beaucoup de sportifs mais il ne les a pas rencontrés. C'est la fluctuation d'échantillonnage.
- M. le maire s'est trompé.

Comme la première explication est peu probable, on préfère la seconde.

III/ Comparer les deux réponses

Dans les deux cas M. le maire affirme que 55 % de ses administrés majeurs pratiquent un sport.

Dans les deux cas, le sondage a trouvé 50 % de sportifs.

Avec le sondage sur 100 personnes l'intervalle de fluctuation est [0,45 ; 0,65].
Comme le sondage a donné 0,50, on accepte l'affirmation de M. le maire.

Avec le sondage sur 1 000 personnes l'intervalle de fluctuation est [0,52 ; 0,58].
Comme le sondage a donné 0,50, on refuse l'affirmation de M. le maire.

Avec un petit effectif, on se contente d'une mauvaise précision, ce qui conduit parfois à accepter une affirmation fausse.

Avec un plus grand effectif, la précision du sondage est meilleure et on s'aperçoit que M. le maire s'est trompé.

François Tuaillon. Mars 2020

Remarques sur les écarts

Avec le sondage sur 1 000 personnes on accepte une fréquence comprise entre 52 % et 58 %. L'écart accepté par rapport à 55 % est de 3 %. On dit « 3 points » car ce n'est pas 3 % de 55 %.

Les sondages utilisent habituellement un effectif d'environ 1 000 personnes. Les sondages ont donc habituellement une précision de 3 points.

Attention. Cette erreur est uniquement due à la fluctuation d'échantillonnage. Il faut aussi tenir compte des autres causes d'erreur : échantillon non représentatif, réponses non sincères etc.

Avec un sondage sur 10 000 personnes, on trouve [0,54 ; 0,56], ce qui correspond à une fréquence comprise entre 54 % et 56 %. L'écart par rapport à 55 % est 1 point.

Avec le sondage sur 100 personnes, on trouve une fréquence comprise 45 % et 65 %. L'écart accepté par rapport à 55 % est 10 points.

Quand l'effectif est multiplié par 10, la précision est environ 3 fois meilleure.

Quand l'effectif est multiplié par 100, la précision est 10 fois meilleure.

Quand l'effectif est multiplié par n, la précision est $\sqrt{n}$ fois meilleure.

Les sondages précis coûtent cher.

6/ Un exemple de différence significative

On compare l'efficacité de deux médicaments.
Le premier médicament est testé sur 1 000 patients et est efficace pour 870 d'entre eux.
Le deuxième médicament est testé sur 800 patients et est efficace pour 720 d'entre eux.
Au seuil de 95 %, ces résultats permettent-ils d'estimer qu'il y a une différence significative entre l'efficacité de ces médicaments ?

Pour le premier médicament, $n = 1\,000 \geq 30$; $n\,p = 870 \geq 5$ et $n\,(1 - p) = 130 \geq 5$.
Pour le deuxième médicament, $n = 8000 \geq 30$; $n\,p = 720 \geq 5$ et $n\,(1 - p) = 80 \geq 5$.
On peut donc chercher un intervalle de confiance.

Pour le premier médicament, la fréquence d'efficacité est $\frac{870}{1000} = 0{,}87$ donc

$\sqrt{\frac{f(1-f)}{n}} = \sqrt{\frac{0{,}87(1-0{,}87)}{1000}} = \sqrt{0{,}000131} \approx 0{,}0106$ donc $1{,}96\sqrt{\frac{f(1-f)}{n}} \approx 0{,}021$.

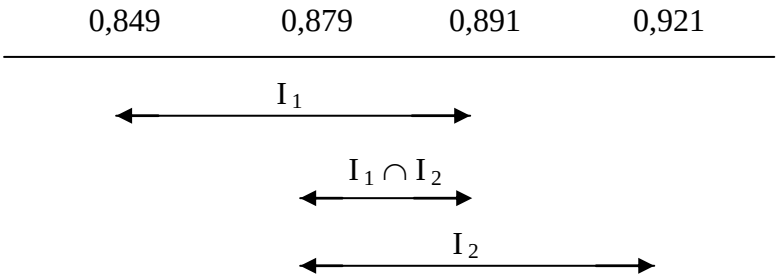
L'intervalle de confiance est donc $I_1 = [\ 0{,}87 - 0{,}021\ ;\ 0{,}87 + 0{,}021\ ] = [\ 0{,}849\ ;\ 0{,}891\ ]$.

Pour le second médicament, la fréquence d'efficacité est $\frac{720}{800} = 0{,}90$ donc

$\sqrt{\frac{f(1-f)}{n}} = \sqrt{\frac{0{,}9(1-0{,}9)}{800}} = \sqrt{0{,}0001125} \approx 0{,}0106$ donc $1{,}96\sqrt{\frac{f(1-f)}{n}} \approx 0{,}021$.

L'intervalle de confiance est donc $I_2 = [\ 0{,}90 - 0{,}021\ ;\ 0{,}90 + 0{,}021\ ] = [\ 0{,}879\ ;\ 0{,}921\ ]$.

François Tuaillon. Mars 2020



$I_1 \cap I_2$ = [0,879 ; 0,891].
Les intervalles de confiance I_1 et I_2 ne sont donc pas disjoints. Au seuil de 95 %, il n'y a donc pas de différence significative d'efficacité entre ces deux médicaments.

François Tuaillon. Mars 2020