# Probabilités et observations économiques

C. Tuleau-Malot

Université de Nice - Sophia Antipolis

# Plan

## variance inconnue

- tout ce qui précède a été fait à variance connue
- et si la variance n'est pas connue?

# variance inconnue

- tout ce qui précède a été fait à variance connue
- et si la variance n'est pas connue?

$\Rightarrow$ on va l'estimer

# construction d'un intervalle de confiance : cadre gaussien et variance inconnue

Quelques définitions :

- Soit $Y_1, \ldots, Y_n$ un $n$-échantillon de loi $\mathcal{N}(0,1)$.
  On considère $C = Y_1^2 + \ldots + Y_n^2$.
  Ainsi définie, on dit que $C$ suit une loi du Chi-Deux $(\chi^2)$ à $n$ degré de liberté.

# construction d'un intervalle de confiance : cadre gaussien et variance inconnue

Quelques définitions :

Soit $Y_1, \ldots, Y_n$ un $n$-échantillon de loi $\mathcal{N}(0, 1)$.
On considère $C = Y_1^2 + \ldots + Y_n^2$.
Ainsi définie, on dit que $C$ suit une loi du Chi-Deux $(\chi^2)$ à $n$ degré de liberté.

- $C$ n peut prendre que des valeurs positives
- le degré de liberté est le nombre de variables $Y_1, \ldots, Y_n$ que l'on peut choisir librement pour entièrement définir $C$.

# construction d'un intervalle de confiance : cadre gaussien et variance inconnue

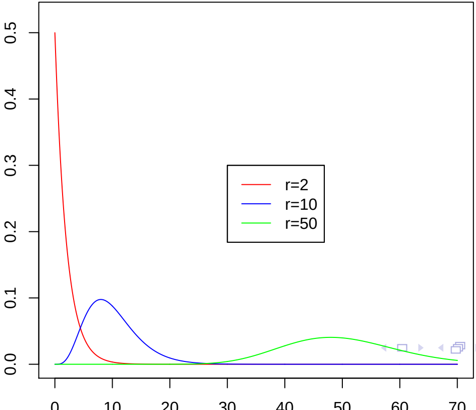
Quelques définitions :

Soit $Y_1, \ldots, Y_n$ un $n$-échantillon de loi $\mathcal{N}(0, 1)$.
On considère $C = Y_1^2 + \ldots + Y_n^2$.
Ainsi définie, on dit que $C$ suit une loi du Chi-Deux $(\chi^2)$ à $n$ degré de liberté.

- $C$ n peut prendre que des valeurs positives
- le degré de liberté est le nombre de variables $Y_1, \ldots, Y_n$ que l'on peut choisir librement pour entièrement définir $C$.

# construction d'un intervalle de confiance : cadre gaussien et variance inconnue

Quelques définitions : Visualisation de la fonction de densité d'une loi du $\chi^2$ selon la valeur du degré de liberté



# construction d'un intervalle de confiance : cadre gaussien et variance inconnue

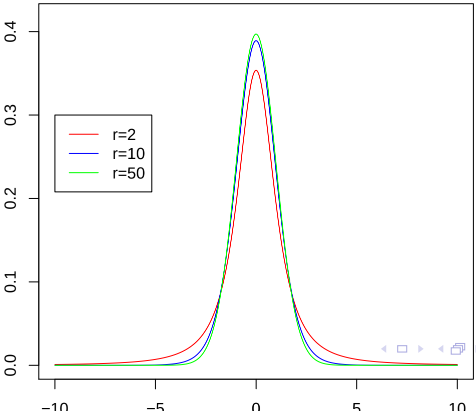
Quelques définitions :
Soit $Y$ une loi $\mathcal{N}(0,1)$ et $V$ une loi du $\chi^2$ à $r$ degré de liberté. On suppose de plus que $Y$ et $V$ sont indépendantes.

On considère $T = \frac{Y}{\sqrt{V/r}}$. Alors ainsi définie, $T$ suit une loi de Student à $r$ degré de liberté.

- la fonction de densité d'une loi de Student est symétrique par rapport à l'axe des ordonnées
- on a ainsi les mêmes propriétés de symétries que pour la loi normale

# construction d'un intervalle de confiance : cadre gaussien et variance inconnue

Quelques définitions : Visualisation de la fonction de densité d'une loi de Student en fonction de son degré de liberté



# construction d'un intervalle de confiance : cadre gaussien et variance inconnue

Quelques résultats :
Soit $X_1, \ldots, X_n$ un $n$-échantillon de loi $\mathcal{N}(\mu, \sigma^2)$ avec $\mu$ et $\sigma^2$ inconnus.

- Soit $U = \frac{n-1}{\sigma^2}\bar{S}_n^2$ avec
  $\bar{S}_n^2 = \frac{1}{n-1}\left((X_1 - \bar{X}_n)^2 + \ldots + (X_n - \bar{X}_n)^2\right)$
  On peut montrer et on admet que $U$ suit une loi du Chi-Deux $(\chi^2)$ à $(n-1)$ degré de liberté.

# construction d'un intervalle de confiance : cadre gaussien et variance inconnue

Quelques résultats :
Soit $X_1, \ldots, X_n$ un $n$-échantillon de loi $\mathcal{N}(\mu, \sigma^2)$ avec $\mu$ et $\sigma^2$ inconnus.

- Soit $U = \frac{n-1}{\sigma^2} \bar{S}_n^2$ avec
  $\bar{S}_n^2 = \frac{1}{n-1} \left( (X_1 - \bar{X}_n)^2 + \ldots + (X_n - \bar{X}_n)^2 \right)$
  On peut montrer et on admet que $U$ suit une loi du Chi-Deux $(\chi^2)$ à $(n-1)$ degré de liberté.

- On admet aussi que $U$ (ou $\bar{S}_n^2$) et $Z = \sqrt{n} \frac{\bar{X}_n - \mu}{\sigma}$ sont des variables indépendantes

# construction d'un intervalle de confiance : cadre gaussien et variance inconnue

Quelques résultats :

- Si on considère

$$T = \frac{Z}{\sqrt{U/(n-1)}}$$

alors $T$ suit une loi de Student à $(n-1)$ degré de liberté, par définition de la loi de Student.

# construction d'un intervalle de confiance : cadre gaussien et variance inconnue

Quelques résultats :

- Si on considère

$$T = \frac{Z}{\sqrt{U/(n-1)}}$$

alors $T$ suit une loi de Student à $(n-1)$ degré de liberté, par définition de la loi de Student.

On montre, après avoir remplacé $Z$ par sa définition et procédé à des simplifications que :

$$T = \sqrt{n}\frac{\bar{X}_n - \mu}{\bar{S}_n}$$

# construction d'un intervalle de confiance : cadre gaussien et variance inconnue

Quelques résultats :
Si l'on compare les expressions de $Z$ et $T$ on voit que $T$ est $Z$ dans lequel on a remplacé $\sigma$ par $\bar{S}_n$.

Quand on remplace un vrai paramètre par un estimateur, on peut changer la distribution. Par exemple, ici on passe d'une loi normale à une loi de Student.

# construction d'un intervalle de confiance : cadre gaussien et variance inconnue

En menant les calculs comme précédemment, on cherche $\epsilon$ tel que :

$$
\begin{aligned}
P(\mu \notin [\bar{X}_n - \epsilon; \bar{X}_n + \epsilon]) &= P(\sqrt{n}.\frac{-\epsilon}{\bar{S}_n} \geq \sqrt{n}.\frac{\bar{X}_n - \mu}{\bar{S}_n} \\
&\quad \text{ou } \sqrt{n}.\frac{\bar{X}_n - \mu}{\bar{S}_n} \geq \sqrt{n}.\frac{\epsilon}{\bar{S}_n}) \\
&\text{on réduit par } \bar{S}_n/\sqrt{n} \\
&\quad \text{au lieu de réduire par } \sigma/\sqrt{n} \\
&= P(\sqrt{n}.\frac{-\epsilon}{\bar{S}_n} \geq \sqrt{n}.\frac{\bar{X}_n - \mu}{\bar{S}_n}) \\
&\quad + P(\sqrt{n}.\frac{\bar{X}_n - \mu}{\bar{S}_n} \geq \sqrt{n}.\frac{\epsilon}{\bar{S}_n})
\end{aligned}
$$

# construction d'un intervalle de confiance : cadre gaussien et variance inconnue

En menant les calculs comme précédemment, on cherche $\epsilon$ tel que :

$$
\begin{aligned}
P(\mu \notin [\bar{X}_n - \epsilon; \bar{X}_n + \epsilon]) &= P(\sqrt{n}.\frac{-\epsilon}{\bar{S}_n} \geq \sqrt{n}.\frac{\bar{X}_n - \mu}{\bar{S}_n} \\
&\quad \text{ou } \sqrt{n}.\frac{\bar{X}_n - \mu}{\bar{S}_n} \geq \sqrt{n}.\frac{\epsilon}{\bar{S}_n}) \\
&= P(\sqrt{n}.\frac{-\epsilon}{\bar{S}_n} \geq \sqrt{n}.\frac{\bar{X}_n - \mu}{\bar{S}_n}) \\
&\quad + P(\sqrt{n}.\frac{\bar{X}_n - \mu}{\bar{S}_n} \geq \sqrt{n}.\frac{\epsilon}{\bar{S}_n}) \\
&= P(T \leq \sqrt{n}.\frac{-\epsilon}{\bar{S}_n}) + P(T \geq \sqrt{n}.\frac{\epsilon}{\bar{S}_n})
\end{aligned}
$$

# construction d'un intervalle de confiance : cadre gaussien et variance inconnue

En menant les calculs comme précédemment, on cherche $\epsilon$ tel que :

$$\begin{aligned}
P(\mu \notin [\bar{X}_n - \epsilon; \bar{X}_n + \epsilon]) &= P(\sqrt{n}.\frac{-\epsilon}{\bar{S}_n} \geq \sqrt{n}.\frac{\bar{X}_n - \mu}{\bar{S}_n} \\
&\quad \text{ou } \sqrt{n}.\frac{\bar{X}_n - \mu}{\bar{S}_n} \geq \sqrt{n}.\frac{\epsilon}{\bar{S}_n}) \\
&= P(\sqrt{n}.\frac{-\epsilon}{\bar{S}_n} \geq \sqrt{n}.\frac{\bar{X}_n - \mu}{\bar{S}_n}) \\
&\quad + P(\sqrt{n}.\frac{\bar{X}_n - \mu}{\bar{S}_n} \geq \sqrt{n}.\frac{\epsilon}{\bar{S}_n}) \\
&= 2.P(T \geq \sqrt{n}.\frac{\epsilon}{\bar{S}_n})
\end{aligned}$$

dernière égalité : par symétrie de la loi de Student (même phénomène que pour la loi normale)

# construction d'un intervalle de confiance : cadre gaussien et variance inconnue

On lit ainsi dans la table associée $t_{1-\alpha/2,n-1}$ de sorte que :

$$P(T \geq t_{1-\alpha/2,n-1}) = \alpha/2$$

# construction d'un intervalle de confiance : cadre gaussien et variance inconnue

On lit ainsi dans la table associée $t_{\alpha/2,n-1}$ de sorte que :

$$P(T \geq t_{1-\alpha/2,n-1}) = \alpha/2$$

Au final, on obtient comme intervalle :

$$\left[\bar{X}_n - \frac{t_{1-\alpha/2,n-1}.\bar{S}_n}{\sqrt{n}}, \bar{X}_n + \frac{t_{1-\alpha/2,n-1}.\bar{S}_n}{\sqrt{n}}\right]$$

# Exemple

On souhaite déterminer la dépense moyenne des clients d'un centre commercial.

On suppose que la distribution des dépenses des clients suit une loi normale. On considère un échantillon de taille $n = 16$ pour lequel

$\bar{x}_n = 140$ et $\bar{s}_n = 30$.

Déterminer l'intervalle de confiance pour les dépenses moyennes au niveau 90%.

# Exemple : solution

On cherche un intervalle de confiance pour le paramètre d'espérance de la loi.

# Exemple : solution

On cherche un intervalle de confiance pour le paramètre d'espérance de la loi. D'après ce qui précède, on sait que

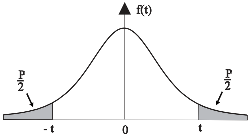
l'intervalle de confiance a pour expression:

$$\left[140 - \frac{30.t_{1-\alpha/2,15}}{4}, 140 + \frac{30.t_{1-\alpha/2;15}}{4}\right]$$

# Exemple : solution

Table de la loi de Student

Valeurs de **T** ayant la probabilité **P** d'être dépassées en valeur absolue




| $\nu$ | P =0,90 | 0,80 | 0,70 | 0,60 | 0,50 | 0,40 | 0,30 | 0,20 | 0,10 | 0,05 | 0,02 | 0,01 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0,158 | 0,325 | 0,510 | 0,727 | 1,000 | 1,376 | 1,963 | 3,078 | 6,314 | 12,706 | 31,821 | 63,657 |
| 2 | 0,142 | 0,289 | 0,445 | 0,617 | 0,816 | 1,061 | 1,386 | 1,886 | 2,920 | 4,303 | 6,965 | 9,925 |
| 3 | 0,137 | 0,277 | 0,424 | 0,584 | 0,765 | 0,978 | 1,250 | 1,638 | 2,353 | 3,182 | 4,541 | 5,841 |
| 4 | 0,134 | 0,271 | 0,414 | 0,569 | 0,741 | 0,941 | 1,190 | 1,533 | 2,132 | 2,776 | 3,747 | 4,604 |
| 5 | 0,132 | 0,267 | 0,408 | 0,559 | 0,727 | 0,920 | 1,156 | 1,476 | 2,015 | 2,571 | 3,365 | 4,032 |
| 6 | 0,131 | 0,265 | 0,404 | 0,553 | 0,718 | 0,906 | 1,134 | 1,440 | 1,943 | 2,447 | 3,143 | 3,707 |
| 7 | 0,130 | 0,263 | 0,402 | 0,549 | 0,711 | 0,896 | 1,119 | 1,415 | 1,895 | 2,365 | 2,998 | 3,499 |
| 8 | 0,130 | 0,262 | 0,399 | 0,546 | 0,706 | 0,889 | 1,108 | 1,397 | 1,860 | 2,306 | 2,896 | 3,355 |
| 9 | 0,129 | 0,261 | 0,398 | 0,543 | 0,703 | 0,883 | 1,100 | 1,383 | 1,833 | 2,262 | 2,821 | 3,250 |
| 10 | 0,129 | 0,260 | 0,397 | 0,542 | 0,700 | 0,879 | 1,093 | 1,372 | 1,812 | 2,228 | 2,764 | 3,169 |
| 11 | 0,129 | 0,260 | 0,396 | 0,540 | 0,697 | 0,876 | 1,088 | 1,363 | 1,796 | 2,201 | 2,718 | 3,106 |
| 12 | 0,128 | 0,260 | 0,395 | 0,539 | 0,695 | 0,873 | 1,083 | 1,356 | 1,782 | 2,179 | 2,681 | 3,055 |
| 13 | 0,128 | 0,259 | 0,394 | 0,538 | 0,694 | 0,870 | 1,079 | 1,350 | 1,771 | 2,160 | 2,650 | 3,012 |
| 14 | 0,128 | 0,258 | 0,393 | 0,537 | 0,692 | 0,868 | 1,076 | 1,345 | 1,761 | 2,145 | 2,624 | 2,977 |
| 15 | 0,128 | 0,258 | 0,393 | 0,536 | 0,691 | 0,866 | 1,074 | 1,341 | 1,753 | 2,131 | 2,602 | 2,947 |
| 16 | 0,128 | 0,258 | 0,392 | 0,535 | 0,690 | 0,865 | 1,071 | 1,337 | 1,746 | 2,120 | 2,583 | 2,921 |

# Exemple : solution

D'où :

$$\left[140 - \frac{30.1,753}{4}, 140 + \frac{30.1,753}{4}\right]$$

Soit :

$$[126,85; 153,15]$$

# construction d'un intervalle de confiance : cadre non gaussien et variance inconnue

- On veut utiliser le Théorème Central Limit
- On veut utiliser un estimateur de $\sigma^2$

# construction d'un intervalle de confiance : cadre non gaussien et variance inconnue

- On veut utiliser le Théorème Central Limit
- On veut utiliser un estimateur de $\sigma^2$

  Oui mais, on ne peut plus utiliser les propriétés précédentes sur $U$ car nous avions que $U$ suit une loi du chi-deux à $n-1$ degré de liberté, uniquement si l'on sait que les variables sont gaussiennes au départ!

# construction d'un intervalle de confiance : cadre non gaussien et variance inconnue

Dans ce cadre, on applique la méthodologie à variance connue et on substitue à $\sigma$, $\bar{S}_n$ sans modifier la distribution gaussienne.

Autrement dit, on va dire que $\sqrt{n}\frac{\bar{X}_n-\mu}{\bar{S}_n}$ suit approximativement une loi gaussienne centrée réduite (d'espérance 0 et de variance 1).

# construction d'un intervalle de confiance : cadre non gaussien et variance inconnue

Dans ce cadre, on applique la méthodologie à variance connue et on substitue à $\sigma$, $\bar{S}_n$ sans modifier la distribution gaussienne.

Autrement dit, on va dire que $\sqrt{n}\frac{\bar{X}_n-\mu}{\bar{S}_n}$ suit approximativement une loi gaussienne centrée réduite (d'espérance 0 et de variance 1).

Cela est possible grâce au lemme de Slutsky!

# construction d'un intervalle de confiance : cadre non gaussien et variance inconnue

Dans ce cadre, on applique la méthodologie à variance connue et on substitue à $\sigma$, $\bar{S}_n$ sans modifier la distribution gaussienne.

Autrement dit, on va dire que $\sqrt{n}\frac{\bar{X}_n-\mu}{\bar{S}_n}$ suit approximativement une loi gaussienne centrée réduite (d'espérance 0 et de variance 1).

Cela est possible grâce au lemme de Slutsky! (mais il faut que $n$ soit suffisamment grand!)

Dans ce cas, l'intervalle est alors :

$$\left[\bar{X}_n - \frac{z_{1-\alpha/2}.\bar{S}_n}{\sqrt{n}}, \bar{X}_n + \frac{z_{1-\alpha/2}.\bar{S}_n}{\sqrt{n}}\right]$$

# Résumé

Intervalle de confiance pour l'espérance :

- Cadre gaussien et variance connue :
  $\left[\bar{X}_n - \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2}; \bar{X}_n + \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2}\right]$
- Cadre gaussien et variance inconnue :
  $\left[\bar{X}_n - \frac{\bar{S}_n}{\sqrt{n}} t_{1-\alpha/2;n-1}; \bar{X}_n + \frac{\bar{S}_n}{\sqrt{n}} t_{1-\alpha/2;n-1}\right]$
- Cadre non gaussien et variance connue :
  $\left[\bar{X}_n - \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2}; \bar{X}_n + \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2}\right]$ ($n$ suffisamment grand)
- Cadre non gaussien et variance inconnue :
  $\left[\bar{X}_n - \frac{\bar{S}_n}{\sqrt{n}} z_{1-\alpha/2}; \bar{X}_n + \frac{\bar{S}_n}{\sqrt{n}} z_{1-\alpha/2}\right]$ ($n$ suffisamment grand)

# Intervalle de confiance pour une proportion

Un enseignant souhaite savoir si son enseignement est satisfaisant pour les étudiants. Il interroge les étudiants de la promotion actuelle. Sur les 250 étudiants interrogés, 82 ont déclaré ne pas être satisfaits. Construire un intervalle de confiance de niveau 95% de la proportion d'étudiants insatisfaits.

# Intervalle de confiance pour une proportion

Un enseignant souhaite savoir si son enseignement est satisfaisant pour les étudiants. Il interroge les étudiants de la promotion actuelle. Sur les 250 étudiants interrogés, 82 ont déclaré ne pas être satisfaits. Construire un intervalle de confiance de niveau 95% de la proportion d'étudiants insatisfaits.

Contexte :

- on considère une proportion $p$ d'un caractère A (proportion sur la population

# Intervalle de confiance pour une proportion

Un enseignant souhaite savoir si son enseignement est satisfaisant pour les étudiants. Il interroge les étudiants de la promotion actuelle. Sur les 250 étudiants interrogés, 82 ont déclaré ne pas être satisfaits. Construire un intervalle de confiance de niveau 95% de la proportion d'étudiants insatisfaits.

Contexte :

- on considère une proportion $p$ d'un caractère A (proportion sur la population
- on a les observations $X_1, \ldots, X_n$ avec $X_i = 1$ si on a le caractère A pour l'individu $i$ et $X_i = 0$ sinon.

# Intervalle de confiance pour une proportion

Un enseignant souhaite savoir si son enseignement est satisfaisant pour les étudiants. Il interroge les étudiants de la promotion actuelle. Sur les 250 étudiants interrogés, 82 ont déclaré ne pas être satisfaits. Construire un intervalle de confiance de niveau 95% de la proportion d'étudiants insatisfaits.

Contexte :

- on considère une proportion $p$ d'un caractère A (proportion sur la population
- on a les observations $X_1, \ldots, X_n$ avec $X_i = 1$ si on a le caractère A pour l'individu $i$ et $X_i = 0$ sinon.
- modélisation de $X_i$ : loi de Bernoulli de paramètre $p$

# Intervalle de confiance pour une proportion

Un enseignant souhaite savoir si son enseignement est satisfaisant pour les étudiants. Il interroge les étudiants de la promotion actuelle. Sur les 250 étudiants interrogés, 82 ont déclaré ne pas être satisfaits. Construire un intervalle de confiance de niveau 95% de la proportion d'étudiants insatisfaits.

Contexte :

- on considère une proportion $p$ d'un caractère A (proportion sur la population
- on a les observations $X_1, \ldots, X_n$ avec $X_i = 1$ si on a le caractère A pour l'individu $i$ et $X_i = 0$ sinon.
- modélisation de $X_i$ : loi de Bernoulli de paramètre $p$
- estimateur de $p$ : $\hat{p}_n = \frac{X_1 + \ldots + X_n}{n}$ (car $\mathbb{E}[X_i] = p$)

# Intervalle de confiance pour une proportion

- Même contexte que précédemment

$\Rightarrow$ on peut appliquer ce qui précède

# Intervalle de confiance pour une proportion

- Même contexte que précédemment

  $\Rightarrow$ on peut appliquer ce qui précède

- cas fréquent et plus simple

  $\Rightarrow$ faire une étude à part

# Intervalle de confiance pour une proportion

Distribution de $n.\hat{p}_n$ ?

$$n.\hat{p}_n = X_1 + \ldots + X_n$$

# Intervalle de confiance pour une proportion

Distribution de $\hat{p}_n$ ?

$$n.\hat{p}_n = X_1 + \ldots + X_n$$

Somme de variables de Bernoulli de paramètre $p$ et indépendantes

$\Rightarrow$ $n.\hat{p}_n$ suit une loi binomiale de paramètre $n$ et $p$

# Intervalle de confiance pour une proportion

- Il n'est pas facile de calculer un intervalle de confiance à partir de la loi binomiale

# Intervalle de confiance pour une proportion

- Il n'est pas facile de calculer un intervalle de confiance à partir de la loi binomiale
- $\hat{p}_n$ est une moyenne empirique

# Intervalle de confiance pour une proportion

- Il n'est pas facile de calculer un intervalle de confiance à partir de la loi binomiale

- $\hat{p}_n$ est une moyenne empirique

  $\Rightarrow$ utilisation du théorème central limit

# Intervalle de confiance pour une proportion

- Il n'est pas facile de calculer un intervalle de confiance à partir de la loi binomiale
- $\hat{p}_n$ est une moyenne empirique

$\Rightarrow$ utilisation du théorème central limit

$\Rightarrow$ $\hat{p}_n$ suit approximativement une loi $\mathcal{N}(p, \frac{p(1-p)}{n})$

# Intervalle de confiance pour une proportion

La variance est ici inconnue mais dans le cas présent, on va faire dans un premier temps comme si on connaissait la variance.

# Intervalle de confiance pour une proportion

La variance est ici inconnue mais dans le cas présent, on va faire dans un premier temps comme si on connaissait la variance.

$$\left[\hat{p}_n - \frac{\sqrt{p(1-p)}}{\sqrt{n}} z_{1-\alpha/2} ; \hat{p}_n - \frac{\sqrt{p(1-p)}}{\sqrt{n}} z_{1-\alpha/2}\right]$$

# Intervalle de confiance pour une proportion

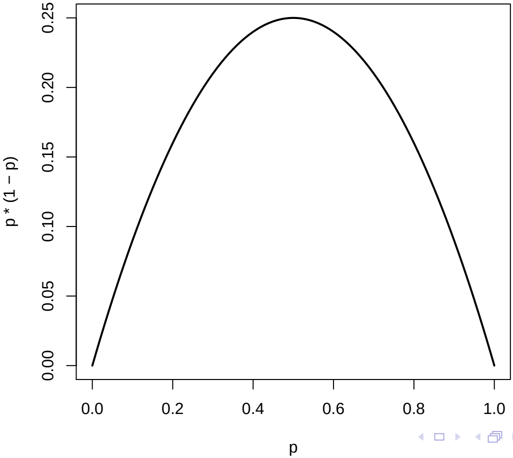
La variance est ici inconnue mais dans le cas présent, on va faire dans un premier temps comme si on connaissait la variance.

$$\left[\hat{p}_n - \frac{\sqrt{p(1-p)}}{\sqrt{n}} z_{1-\alpha/2}; \hat{p}_n + \frac{\sqrt{p(1-p)}}{\sqrt{n}} z_{1-\alpha/2}\right]$$

Cela ne peut pas être un intervalle de confiance! (en effet il dépend de $p$!)

# Intervalle de confiance pour une proportion

Solution 1 : graphe de la fonction $p \to p(1-p)$



# Intervalle de confiance pour une proportion

Solution 1 :
On voit que la fonction $p \to p(1-p)$ est croissante sur $[0; 1/2]$ puis décroissante sur $[1/2; 1]$.
Donc elle est maximale en $1/2$ où elle vaut $1/4$.

Autrement dit :

$$\text{pour tout p dans [0;1]} \quad p.(1-p) \leq 0.25 = \frac{1}{4}$$

# Intervalle de confiance pour une proportion

Solution 1 :

On voit que la fonction $p \to p(1-p)$ est croissante sur $[0;1/2]$ puis décroissante sur $[1/2;1]$.
Donc elle est maximale en $1/2$ où elle vaut $1/4$.

Autrement dit :

$$\text{pour tout p dans [0;1]} \quad p.(1-p) \leq 0.25 = \frac{1}{4}$$

Et donc un intervalle de confiance est :

$$\left[\hat{p}_n - \frac{1}{\sqrt{4.n}} z_{1-\alpha/2} ; \hat{p}_n + \frac{1}{\sqrt{4.n}} z_{1-\alpha/2}\right]$$

On a majoré $p(1-p)$ par $1/4$.

# Intervalle de confiance pour une proportion

Solution 2 :
On remplace dans la formule de la variance $p$ par son estimateur sans changer la loi. C'est le lemme de Slutsky qui le permet. Il faut pour pouvoir opérer cela que le nombre d'observations $n$ soit suffisamment grand. Et donc, en procédant comme ce qui a pu

être fait avant, un intervalle de confiance est :

$$\left[\hat{p}_n - \frac{\sqrt{\hat{p}_n(1-\hat{p}_n)}}{\sqrt{n}} z_{1-\alpha/2} ; \hat{p}_n + \frac{\sqrt{\hat{p}_n(1-\hat{p}_n)}}{\sqrt{n}} z_{1-\alpha/2}\right]$$

# Taille de l'échantillon

On peut aborder le problème autrement en évaluant la taille que doit avoir l'échantillon pour avoir une marge d'erreur fixée.

# Taille de l'échantillon

On peut aborder le problème autrement en évaluant la taille que doit avoir l'échantillon pour avoir une marge d'erreur fixée.

Exemple :

Un magasin veut procéder à une enquête de satisfaction de sa clientèle. Le magasin aimerait connaître le taux de satisfaction avec une précision de $\pm$ 0,03.

Combien de clients doit-on interroger?

# Taille de l'échantillon

- On s'intéresse à un taux de satisfaction

$\Rightarrow$ cadre de la proportion

# Taille de l'échantillon

- On s'intéresse à un taux de satisfaction

  $\Rightarrow$ cadre de la proportion

- un intervalle de confiance est donc :

$$\left[\hat{p}_n - \frac{\sqrt{\hat{p}_n(1-\hat{p}_n)}}{\sqrt{n}} z_{1-\alpha/2} ; \hat{p}_n + \frac{\sqrt{\hat{p}_n(1-\hat{p}_n)}}{\sqrt{n}} z_{1-\alpha/2}\right]$$

# Taille de l'échantillon

- On s'intéresse à un taux de satisfaction

  $\Rightarrow$ cadre de la proportion

- un intervalle de confiance est donc :

$$\left[\hat{p}_n - \frac{\sqrt{\hat{p}_n(1-\hat{p}_n)}}{\sqrt{n}} z_{1-\alpha/2} ; \hat{p}_n + \frac{\sqrt{\hat{p}_n(1-\hat{p}_n)}}{\sqrt{n}} z_{1-\alpha/2}\right]$$

- précision de $\pm$ 0.03

$$\Rightarrow \frac{\sqrt{\hat{p}_n(1-\hat{p}_n)}}{\sqrt{n}} z_{1-\alpha/2} \leq 0.03$$

# Taille de l'échantillon

$$\frac{\sqrt{\hat{p}_n(1-\hat{p}_n)}}{\sqrt{n}} z_{1-\alpha/2} \leq 0.03$$

équivalent à :

$$\frac{\hat{p}_n(1-\hat{p}_n).z^2_{1-\alpha/2}}{0.03^2} \leq n$$

# Taille de l'échantillon

$$\frac{\sqrt{\hat{p}_n(1-\hat{p}_n)}}{\sqrt{n}} z_{1-\alpha/2} \leq 0.03$$

équivalent à :

$$\frac{\hat{p}_n(1-\hat{p}_n).z^2_{1-\alpha/2}}{0.03^2} \leq n$$

Mais on ne connaît pas $\hat{p}_n$!
$\Rightarrow$ on majore $\hat{p}_n(1-\hat{p}_n)$ par $1/4$

# Taille de l'échantillon

$$\frac{\sqrt{\hat{p}_n(1-\hat{p}_n)}}{\sqrt{n}} z_{1-\alpha/2} \leq 0.03$$

équivalent à :

$$\frac{\hat{p}_n(1-\hat{p}_n).z^2_{1-\alpha/2}}{0.03^2} \leq n$$

Mais on ne connaît pas $\hat{p}_n$!
$\Rightarrow$ on majore $\hat{p}_n(1-\hat{p}_n)$ par $1/4$

$$\frac{z^2_{1-\alpha/2}}{4.0.03^2} \leq n$$

# Introduction

Exemples :

- En 2008, on sait que la proportion de fumeurs était de 38%. Une enquête sur 235 personnes donne 79 fumeurs. Peut-on affirmer que la proportion de fumeurs a baissé?

# Introduction

Exemples :

- En 2008, on sait que la proportion de fumeurs était de 38%. Une enquête sur 235 personnes donne 79 fumeurs. Peut-on affirmer que la proportion de fumeurs a baissé?

- Une chaîne de production remplit des bouteilles d'eau de 1L. On prélève 40 bouteilles dont on mesure la contenance avec précision. On veut voir si la machine est bien réglée, à savoir si la moyenne de remplissage est bien de 1L.

# Formulation mathématique

Formalisons l'exemple 1

- on a des variables de Bernoulli indépendantes de paramètre $p$ inconnu.
- on veut savoir si $p < 0.38$ ou si $p \geq 0.38$.

# Formulation mathématique

Formalisons l'exemple 1

- on a des variables de Bernoulli indépendantes de paramètre $p$ inconnu.
- on veut savoir si $p < 0.38$ ou si $p \geq 0.38$.
- on va vouloir réaliser un test entre :

$$\mathcal{H}_0 : \ p \geq 0.38$$
$$\mathcal{H}_a : \ p < 0.38$$

# Formulation mathématique

Formalisons l'exemple 1

- on a des variables de Bernoulli indépendantes de paramètre $p$ inconnu.
- on veut savoir si $p < 0.38$ ou si $p \geq 0.38$.
- on va vouloir réaliser un test entre :

$$\mathcal{H}_0 : p \geq 0.38$$
$$\mathcal{H}_a : p < 0.38$$

$\mathcal{H}_0$ : hypothèse nulle et $\mathcal{H}_a$ : hypothèse alternative

# Principe d'un test

Objectif :

à partir des observations d'un échantillon de taille $n$, décider entre les deux hypothèses!

# Théorie

Attention, les deux hypothèses ne jouent pas le même rôle!
Pourquoi?

# Théorie

Attention, les deux hypothèses ne jouent pas le même rôle!

Pourquoi?

décision du test

| | | hypothèse nulle | Hypothèse alternative |
|---|---|---|---|
| réalité | hypothèse nulle | Correct | Incorrect |
| | Hypothèse alternative | Incorrect | Correct |

# Théorie

Deux sources d'erreur :

- décider $\mathcal{H}_a$ alors que c'est $\mathcal{H}_0$ qui est vraie

  $\Rightarrow$ erreur de 1ère espèce

# Théorie

Deux sources d'erreur :

- décider $\mathcal{H}_a$ alors que c'est $\mathcal{H}_0$ qui est vraie
  $\Rightarrow$ erreur de 1ère espèce

- décider $\mathcal{H}_0$ alors que c'est $\mathcal{H}_a$ qui est vraie
  $\Rightarrow$ erreur de 2ème espèce

# Théorie

Deux sources d'erreur :

- décider $\mathcal{H}_a$ alors que c'est $\mathcal{H}_0$ qui est vraie

  $\Rightarrow$ erreur de 1ère espèce

- décider $\mathcal{H}_0$ alors que c'est $\mathcal{H}_a$ qui est vraie

  $\Rightarrow$ erreur de 2ème espèce

On aimerait minimiser ces deux erreurs, mais impossible!

$\Rightarrow$ on choisit de contrôler l'erreur de 1ère espèce

# Théorie

Concrètement, cela signifie que :

- on donne le bénéfice du doute à $\mathcal{H}_0$
- on ne va rejeter $\mathcal{H}_0$ que si les observations penchent fortement en faveur de $\mathcal{H}_a$

# Théorie

Concrètement, cela signifie que :

- on donne le bénéfice du doute à $\mathcal{H}_0$
- on ne va rejeter $\mathcal{H}_0$ que si les observations penchent fortement en faveur de $\mathcal{H}_a$

L'hypothèse alternative est très souvent ce que l'on veut montrer

# Exemples de test

Il existe différents tests :

- Tests d'hypothèse : on va comparer la moyenne sur la population à une valeur de référence

# Exemples de test

Il existe différents tests :

- Tests d'hypothèse : on va comparer la moyenne sur la population à une valeur de référence

- Tests de comparaison : on va comparer la moyenne de deux populations

# Exemples de test

Il existe différents tests :

- Tests d'hypothèse : on va comparer la moyenne sur la population à une valeur de référence
- Tests de comparaison : on va comparer la moyenne de deux populations
- Test d'adéquation : on va regarder si l'on peut considérer que notre variable suit une distribution pré-définie

# Exemples de test

Il existe différents tests :

- Tests d'hypothèse : on va comparer la moyenne sur la population à une valeur de référence

- Tests de comparaison : on va comparer la moyenne de deux populations

- Test d'adéquation : on va regarder si l'on peut considérer que notre variable suit une distribution pré-définie

- Test d'indépendance : on va voir si deux variables aléatoires sont indépendantes ou non

# Stratégie

Dans chacun des cas cités précédemment, on va :

1. Définir les hypothèses $\mathcal{H}_0$ et $\mathcal{H}_a$

# Stratégie

Dans chacun des cas cités précédemment, on va :

1. Définir les hypothèses $\mathcal{H}_0$ et $\mathcal{H}_a$
2. on va calculer une statistique de test (une variable aléatoire) que l'on peut entièrement évaluer à partir d'observations et dont on connaît la loi sous l'hypothèse $\mathcal{H}_0$

# Stratégie

Dans chacun des cas cités précédemment, on va :

1. Définir les hypothèses $\mathcal{H}_0$ et $\mathcal{H}_a$
2. on va calculer une statistique de test (une variable aléatoire) que l'on peut entièrement évaluer à partir d'observations et dont on connaît la loi sous l'hypothèse $\mathcal{H}_0$
3. si la statistique de test, calculée sur l'échantillon, prend une valeur trop improbable, alors on rejette $\mathcal{H}_0$