# Chapitre 4

# DISTRIBUTIONS D'ECHANTILLONNAGE et INTERVALLES DE VARIATION

# Chapitre 4 *(suite 1)*

1. Introduction

2. Estimation d'une moyenne
   - Distribution d'échantillonnage
   - Intervalles de variation

3. Estimation d'une proportion
   - Distribution d'échantillonnage
   - Intervalles de variation

4. Estimation d'une variance
   - Distribution d'échantillonnage
   - Estimation sans biais

# Chapitre 4 *(suite 1)*

## 3. Estimation d'une proportion

**3.1 Variable qualitative**

**3.2 Fréquence empirique : statistique et estimateur**

**3.3 Fluctuations d'échantillonnage de la fréquence empirique**

exemple théorique

**3.4 Propriétés de la fréquence empirique**

**3.5 Distribution d'échantillonnage de la fréquence empirique**

variable qualitative dichotomique :
théorème central-limite

**3.6 Quantiles de la fréquence empirique**

**3.7 Intervalles de variation de la fréquence empirique**

interprétation d'un intervalle de variation

**3.8 Précision dans l'estimation de la proportion**

**3.9 Taille minimum de l'échantillon pour une précision minimum**

# 3. Estimation d'une proportion

**X variable qualitative dichotomique de $\mathcal{P}$ dans E = {oui, non}**

$\Rightarrow$un paramètre

**proportion de *"oui"* = *p*** ***connue*** **dans $\mathcal{P}$**

échantillon de **X** issu de $\mathcal{P}$ de taille ***n***

- étudier les propriétés de l'estimation ponctuelle du paramètre proportion ***p*** par la fréquence ***f*** observée sur un échantillon de taille ***n***

- la fréquence observée ***f*** ***varie*** d'un échantillon à l'autre

- étudier les variations (fluctuations) de la fréquence observée ***f*** sur tous les échantillons de même taille ***n***

# 3.1 Variable qualitative (1)

- **Exemple : résultats au DEUG en 1994**

*$\mathcal{P}$ = { étudiants inscrits en 2d année de DEUG S.H. à Nanterre en 1994}*
***N** = 617*
***X** = "réussite au DEUG" : oui, non*
*variable qualitative dichotomique*
***p** = proportion de réussite = **0,75** **connue** dans $\mathcal{P}$*

*échantillon de **X** issu de $\mathcal{P}$ de taille **n** = 40*

➢ *la proportion de réussite **p** est estimée par la fréquence de réussite observée **f***

① *quelle fréquence **f** s'attend-on à observer sur l'échantillon de taille **n** = 40 ?*

② *quelle précision a-t-on en estimant la proportion de réussite **p** par **f** sur un échantillon de taille **n** = 40 ?*

③ *serait-on surpris d'observer une fréquence de réussite **f** = **88%** sur l'échantillon de taille **n** = 40 ?*

# 3.1 Variable qualitative (2)

- **Exemple : résultats au DEUG en 1994**

*$\mathcal{P}$ = { étudiants inscrits en 2d année de DEUG S.H. à Nanterre en 1994}*
***N** = 617*
***X** = "réussite au DEUG" : oui, non variable qualitative dichotomique*
***p** = proportion de réussite = **0,75 connue** dans $\mathcal{P}$*

*échantillon de **X** issu de $\mathcal{P}$ de taille **n** = 40*

➢ *la proportion de réussite **p** est estimé par la fréquence de réussite observée **f***

④ *à partir de quelle limite une fréquence de réussite **f** observée sur un échantillon de taille **n** = 40 peut-elle être considérée comme "surprenante" ?*

⑤ *dans quel intervalle s'attend-on raisonnablement à observer la fréquence de réussite **f** sur l'échantillon de taille **n** = 40 ?*

# 3.2 Fréquence empirique : statistique et estimateur

La variable qui représente toutes les valeurs observées ***f*** sur tous les échantillons possibles de taille ***n*** est appelée **fréquence empirique** et notée $F_n$

$F_n$ est une **statistique** : une variable calculée à partir des observations $(x_1, x_2, \ldots, x_n)$ qui permet de résumer numériquement ces observations

- les propriétés de la statistique dépendent de la taille de l'échantillon ***n***

***f*** est la valeur calculée sur l'échantillon observé de taille ***n*** : c'est une **estimation** de du paramètre ***p***

- une estimation est propre à l'échantillon observé

$F_n$ est l'ensemble de toutes les estimations possibles ***f*** sur tous les échantillons de taille ***n*** : c'est un **estimateur** de ***p***

# 3.3 Fluctuations d'échantillonnage de la fréquence empirique

population de tous les échantillons de taille ***n***

- un individu est un échantillon de taille ***n***

$F_n$ = *fréquence empirique*

⇨ **variable quantitative** définie de la population des échantillons de taille ***n*** dans **[0, 1]**

⇨ deux paramètres { **moyenne** de $F_n$ ; **écart-type** de $F_n$

⇨ **forme** de la distribution (loi) de $F_n$

➢ **étudier les variations (fluctuations) de la fréquence empirique**

➢ étudier la ***"loi des fréquences"***

# Exemple théorique (1)

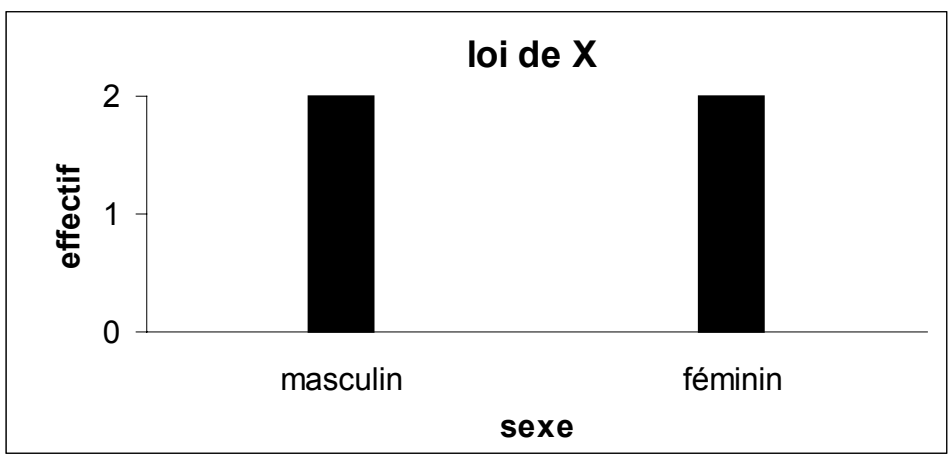
- **Exemple : sexe**

  $\mathcal{P}$ = *{ personnes }* ***N*** = *4*

  ***X*** = *"sexe"*

  *variable qualitative dichotomique sur*
  *E = { féminin, masculin }*

| variable X | $n_i$ | $p_i$ |
|---|---|---|
| masculin = M | 2 | 0,5 |
| féminin = F | 2 | 0,5 |
| total | 4 | 1 |




➔ *un paramètre* **connu** *dans* $\mathcal{P}$

***p*** = *proportion de femmes =* ***0,5***
(**1**–***p*** = *proportion d'hommes =* ***0,5***)

# Exemple théorique (2)

- **Exemple : sexe**

$\mathcal{P}$ = { *personnes* } ***N*** = *4*

***X*** = *"sexe"*

*variable qualitative dichotomique sur E = { féminin, masculin }*

➔ *un paramètre* **connu** *dans* $\mathcal{P}$

***p*** = *proportion de femmes* = ***0,5***

*échantillons de* ***X*** *issu de* $\mathcal{P}$ *de taille* ***n*** = *2 : il y en a* **16**

***pour chacun des 16 échantillons la fréquence observée f varie***

| échantillon | $(X_1, X_2)$ | nb de femmes | $F_2$ |
|---|---|---|---|
| 1 | ($x_1$= M, $x_2$= M) | 0 | 0 |
| 2 | (M, M) | 0 | 0 |
| 3 | (M, F) | 1 | 0,5 |
| 4 | (M, F) | 1 | 0,5 |
| 5 | (M, M) | 0 | 0 |
| 6 | (M, M) | 0 | 0 |
| 7 | (M, F) | 1 | 0,5 |
| 8 | (M, F) | 1 | 0,5 |
| 9 | (F, M) | 1 | 0,5 |
| 10 | (F, M) | 1 | 0,5 |
| 11 | (F, F) | 2 | 1 |

# Exemple théorique (3)

- **Exemple : sexe**

*population des échantillons de taille* ***n*** *= 2* ***de taille 16***

$F_2$ = *"fréquence de femmes"*

*variable quantitative sur [ 0, 1]*

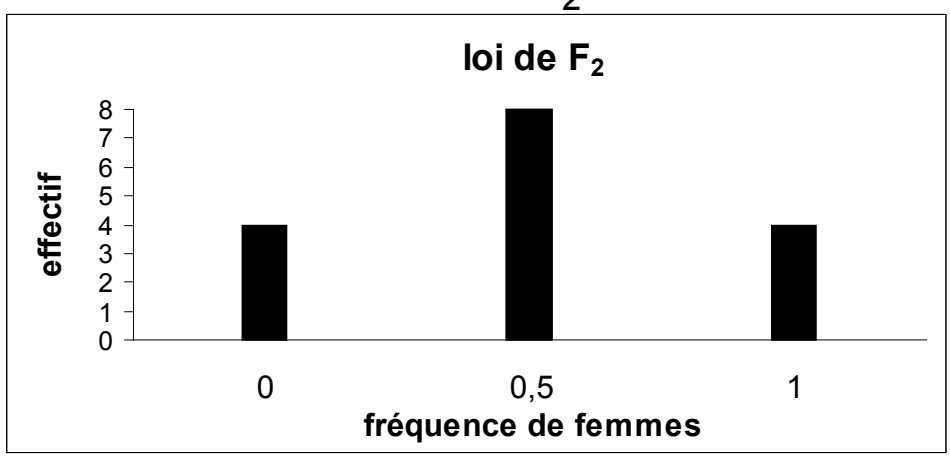
*distribution de* $F_2$ *: "loi des fréquences"*

| variable $F_2$ | $n_i$ | $n_i f_i$ | $n_i f_i^2$ |
|---|---|---|---|
| 0 | 4 | 0 | 0 |
| 0,5 | 8 | 4 | 8×0,25=2 |
| 1 | 4 | 4 | 4 |
| total | 16 | 8 | 6 |

$f$

$\Sigma\, n_i f_i = 8 \qquad \Sigma\, n_i f_i^{\,2} = 6$

➔ *forme de la loi de* $F_2$




➔ *deux paramètres pour la loi de* $F_2$

- *moyenne de* $F_2$ = 8 / 16 = 0,5 = ***p***
- *variance de* $F_2$ = (6 / 16) – $0{,}5^2$ = 0,125 = ***p***(1–***p***) / **2**

# 3.4 Propriétés de la fréquence empirique (1)

- La moyenne de la fréquence empirique $F_n$ est égale à $\boldsymbol{p}$

  *"la moyenne des fréquences est égale à la proportion dans la population"*

  $\Rightarrow$la fréquence empirique $F_n$ est un estimateur sans biais de $\boldsymbol{p}$

- La variance de la fréquence empirique $F_n$ est égale à $\dfrac{\boldsymbol{p}(1-\boldsymbol{p})}{\boldsymbol{n}}$
  - plus la taille de l'échantillon est grande plus l'estimation de la proportion est précise
  - la variabilité de l'estimation tend vers 0 quand la taille de l'échantillon devient infinie

  $\Rightarrow$la fréquence empirique $F_n$ est un estimateur convergent de variance minimum de $\boldsymbol{p}$

# 3.4 Propriétés de la fréquence empirique (2)

- **X variable qualitative dichotomique définie de $\mathcal{P}$ dans E = { oui, non }**

  **proportion de "oui" = $p$**

  **échantillon de X issu de $\mathcal{P}$ de taille $n$**

- **$F_n$ fréquence empirique sur les échantillons de taille $n$**
  **variable quantitative sur [0, 1]**

$$\begin{cases} \text{moyenne de } F_n = p \\ \text{variance de } F_n = p(1-p)/n \\ \text{écart-type de } F_n = \sqrt{p(1-p)/n} \end{cases}$$

$\Rightarrow$ la fréquence empirique est un ***bon*** estimateur de la proportion $p$ de X dans $\mathcal{P}$ : il est sans biais, convergent et de variance minimum

- en pratique, on utilisera la **fréquence observée** $f$ sur l'échantillon comme **estimation ponctuelle** de $p$

# 3.4 Propriétés de la fréquence empirique (3)

- **Exemple : résultats au DEUG en 1994**

$\mathcal{P}$ = { *étudiants inscrits en 2d année de DEUG S.H. à Nanterre en 1994*}

***N*** = *617*

***X*** = *"réussite au DEUG" : oui, non*
*variable qualitative dichotomique*
*un paramètre* ***connu*** *dans* $\mathcal{P}$

***p*** = *proportion de réussite =* ***0,75***

*échantillon de* ***X*** *issu de* $\mathcal{P}$ *de taille*
***n*** *= 40*

➢ *la moyenne de* $F_n$ = ***p*** = *0,75*

① *on s'attend <u>en moyenne</u> à observer une fréquence de réussite égale à* ***p*** *= 75% sur l'échantillon de taille* ***n*** *= 40, comme sur n'importe quel autre échantillon de taille* ***n***

➢ *l'écart-type de* $F_n = \sqrt{\boldsymbol{p}(1-\boldsymbol{p}) / \boldsymbol{n}}$
$= \sqrt{0,75 \times 0,25/40}$
$= 0,0685$

② *la variabilité de l'estimation de la proportion de réussite* ***p*** *par la fréquence observée sur un échantillon de taille* ***n*** *= 40 est de 0,0685 (6,85%)*

# 3.4 Propriétés de la fréquence empirique (4)

- **Exemple : résultats au DEUG en 1994**

  $\mathcal{P}$ = { *étudiants inscrits en 2d année de DEUG S.H. à Nanterre en 1994*}

  ***N*** = *617*

  ***X*** = *"réussite au DEUG" : oui, non*
  *variable qualitative dichotomique*
  *un paramètre* ***connu*** *dans* $\mathcal{P}$

  ***p*** = *proportion de réussite =* ***0,75***

  *échantillon de* ***X*** *issu de* $\mathcal{P}$ *de taille*
  ***n*** *= 100*

  ➢ *la moyenne de* $F_n$ = ***p*** = *0,75*

  ➢ *l'écart-type de* $F_n = \sqrt{\boldsymbol{p}(1-\boldsymbol{p}) / \boldsymbol{n}}$
  $= \sqrt{0,75 \times 0,25/100}$
  $= 0,0433$

  ② *la variabilité de l'estimation de la proportion de réussite* ***p*** *par la fréquence observée sur un échantillon de taille* ***n*** *= 100 est de 0,0433 (4,33%)*
  *inférieure à celle obtenue pour un échantillon de taille* ***n*** *= 40*

# 3.5 Distribution d'échantillonnage de la fréquence empirique (1)

## Variable qualitative dichotomique : théorème central-limite

- X variable *qualitative dichotomique* de proportion de *"oui"* $p$ dans $\mathcal{P}$
  - si $n$ est "*suffisamment*" grand $n \geq 30$
  - si les proportions $p$ et $(1-p)$ ne sont pas "*trop*" proches de 0 ou de 1
    $np \geq 5$ et $n(1-p) \geq 5$

la fréquence empirique $F_n$ *suit* ***approximativement*** *un modèle normal* de moyenne $p$ et d'écart-type $\sqrt{p(1-p)/n}$

$$F_n \underset{approx}{\sim} \mathcal{N}\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

# 3.5 Distribution d'échantillonnage de la fréquence empirique (2)

## Variable qualitative dichotomique : théorème central-limite

$$F_n \underset{approx}{\sim} \mathcal{N}\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

$\Rightarrow$lorsque les valeurs de $p$ et $n$ sont ***connues*** on peut calculer approximativement à l'aide du modèle normal sur $F_n$

- **la fonction de répartition** $P(F_n \leq f)$ **proportion d'échantillons de taille** *n* **de X issus de** $\mathcal{P}$ **ayant une fréquence observée inférieure à** *f*
- **la loi de probabilité** $P(c \leq F_n \leq d)$
- **les quantiles**
- **les intervalles de variation**

# 3.5 Distribution d'échantillonnage de la fréquence empirique (3)

## Conditions de l'approximation normale

- les proportions $p$ et $(1-p)$ ne sont pas "*trop*" proches de 0 ou de 1

  $np \geq 5$ et $n(1-p) \geq 5$

  - **sur un échantillon de taille *n***
    ***np*** : effectif moyen "*attendu*" de "*oui* "
    ***n*(1–*p*)** : effectif moyen "*attendu*" de "*non*"
  - **plus la taille de l'échantillon est grande plus la proportion peut être "*proche*" de 0 ou de 1**

| proportion | | taille de l'échantillon |
|---|---|---|
| ***p*** | 1–***p*** | ***n*** |
| 0,5 | 0,5 | 30 |
| 0,25 | 0,75 | 30 |
| 1/6 ≈ 0,167 | 5/6 ≈ 0,833 | 30 |
| 0,15 | 0,85 | 34 |
| 0,1 | 0,9 | 50 |
| 0,05 | 0,95 | 100 |
| 0,01 | 0,99 | 500 |
| 0,001 | 0,999 | 5000 |

# Exemple variable qualitative (1)

- **Exemple : résultats au DEUG en 1994**

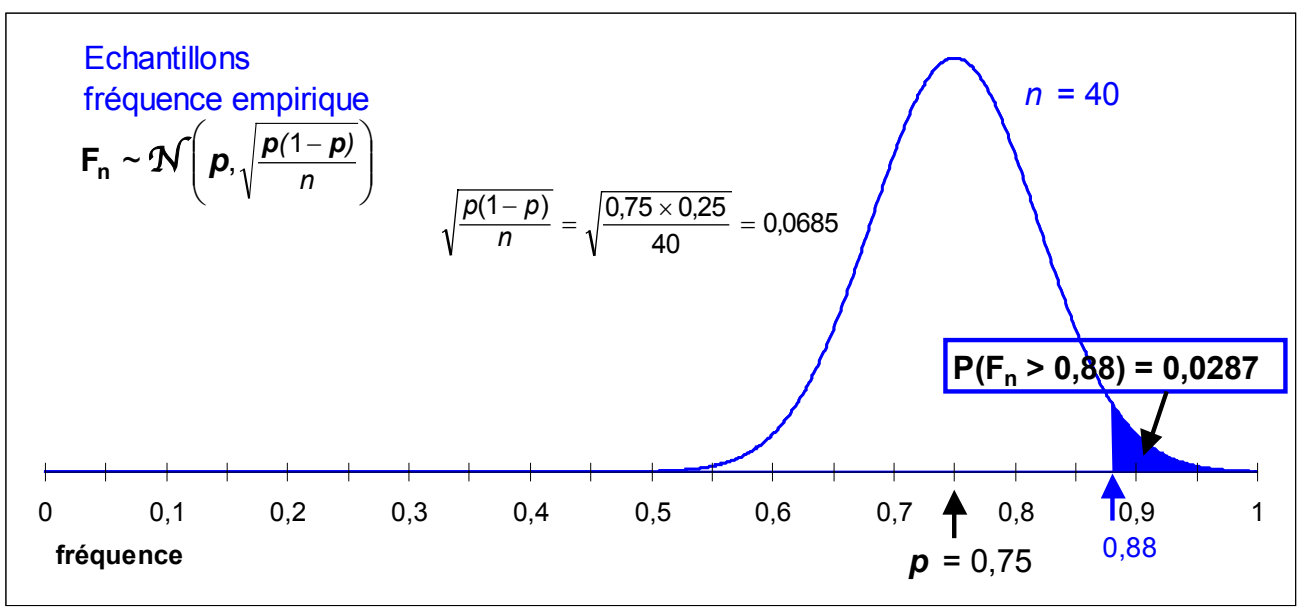
$\mathcal{P}$ = { *étudiants inscrits en 2d année de DEUG S.H. à Nanterre en 1994*}

***X*** = *"réussite au DEUG" : oui, non*
*variable qualitative dichotomique*
*un paramètre* ***connu*** *dans* $\mathcal{P}$
***p*** = *proportion de réussite =* ***0,75***

*échantillon de* ***X*** *issu de* $\mathcal{P}$ *de taille*
$n = 40$ ($n \geq 30$)
$np = 30 \geq 5$ et $n(1-p) = 10 \geq 5$

*alors la fréquence empirique* $F_n$ *suit approximativement une loi normale*

$$F_n \underset{approx}{\sim} \mathcal{N}(0{,}75;0{,}0685)$$



# Exemple variable qualitative (2)

- **Exemple : résultats au DEUG en 1994**

$\mathcal{P}$ = { *étudiants inscrits en 2d année de DEUG S.H. à Nanterre en 1994*}

***X*** = *"réussite au DEUG" : oui, non*
*variable qualitative dichotomique*
*un paramètre* ***connu*** *dans* $\mathcal{P}$
***p*** = *proportion de réussite =* ***0,75***

*échantillon de* ***X*** *issu de* $\mathcal{P}$ *de taille*

***n*** = *40* $\quad F_n \underset{approx}{\sim} \mathcal{N}(0{,}75;0{,}0685)$

➢ $P(F_n > 0{,}88) = 0{,}0287$

➔ *2,87% des échantillons de taille* ***n*** *= 40 ont une fréquence de réussite supérieure à 88%*

③ *on serait surpris d'observer une fréquence de réussite supérieure à 88% sur un échantillon de taille* ***n*** *= 40*

# Exemple variable qualitative (3)

- **Exemple : résultats au DEUG en 1994**

$\mathcal{P}$ = { *étudiants inscrits en 2d année de DEUG S.H. à Nanterre en 1994*}

***X*** = *"réussite au DEUG" : oui, non*
*variable qualitative dichotomique*

*échantillon de* ***X*** *issu de* $\mathcal{P}$ *de taille*

***n*** = *30*

$np = 22,5 \geq 5$ et $n(1-p) = 7,5 \geq 5$

$F_n \underset{approx}{\sim} \mathcal{N}(0,75;0,0791)$

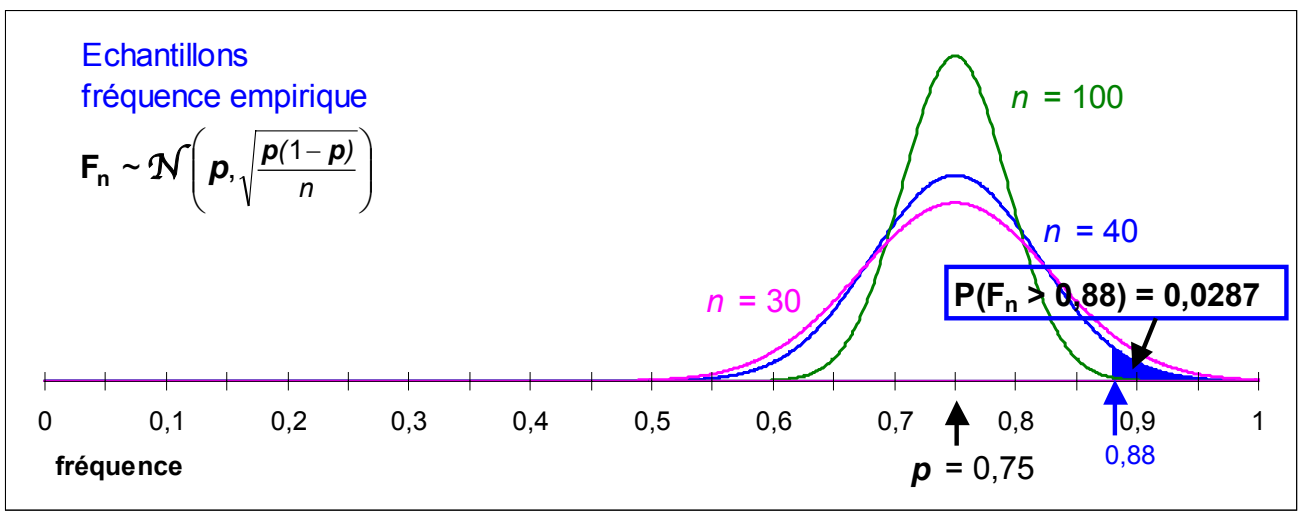
➢ $P(F_n > 0,88) = 0,0505$

***n*** =*100*

$np = 75 \geq 5$ et $n(1-p) = 25 \geq 5$

$F_n \underset{approx}{\sim} \mathcal{N}(0,75;0,0433)$

➢ $P(F_n > 0,88) = 0,00135$



# 3.6 Quantiles de la fréquence empirique (1)

**X variable qualitative dichotomique de $\mathcal{P}$ dans E = { oui, non }**

$\Rightarrow$un paramètre ***connu*** dans $\mathcal{P}$

**proportion de *"oui"* = *p***

échantillon de **X** issu de $\mathcal{P}$ de taille ***n***

- rechercher la limite à partir de laquelle la fréquence observée sur l'échantillon de taille ***n***, estimation de la proportion ***p***, peut être considérée comme *"surprenante"* :
  - soit **"étonnamment" faible** :
    $\alpha$% des fréquences les plus faibles
  - soit **"étonnamment" élevée** :
    $\alpha$% des fréquences les plus élevées

$\Rightarrow$ces limites sont définies à partir des *quantiles* de la fréquence empirique $F_n$

$\Rightarrow$elles sont calculées en utilisant le fait que la fréquence empirique $F_n$ suit approximativement une loi normale

# 3.6 Quantiles de la fréquence empirique (2)

**X** variable **qualitative dichotomique** de $\mathcal{P}$
dans E = { oui, non }
de **proportion de *"oui"*** = $p$ dans $\mathcal{P}$

échantillon de **X** issu de $\mathcal{P}$ de taille $n \geq 30$
avec $np \geq 5$ et $n(1-p) \geq 5$

- la fréquence empirique $F_n$ suit approximativement un modèle normal de moyenne $p$ et d'écart-type $\sqrt{p(1-p)/n}$

$$F_n \underset{approx}{\sim} \mathcal{N}\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

$\Rightarrow$le quantile d'ordre $\alpha$ de $F_n$ s'écrit :

$$p + z_\alpha \sqrt{\frac{p(1-p)}{n}}$$

où $z_\alpha$ quantile d'ordre $\alpha$ de $Z \sim \mathcal{N}(0,1)$

$\Rightarrow$le quantile d'ordre $\alpha$ de $F_n$ avec $\alpha < 0{,}5$ s'écrit :

$$p - z_{1-\alpha} \sqrt{\frac{p(1-p)}{n}}$$

où $z_{1-\alpha}$ quantile d'ordre $1-\alpha$ de $Z \sim \mathcal{N}(0,1)$

# Exemple variable qualitative (1)

- **Exemple : résultats au DEUG en 1994**

$\mathcal{P}$ = { *étudiants inscrits en 2d année de DEUG S.H. à Nanterre en 1994*}

***X*** = *"réussite au DEUG" : oui, non*
*variable qualitative dichotomique*
*un paramètre* ***connu*** *dans* $\mathcal{P}$
***p*** = *proportion de réussite =* ***0,75***

*échantillon de* ***X*** *issu de* $\mathcal{P}$ *de taille*
***n*** = *40 (****n*** $\geq$ *30)*
*np = 30* $\geq$ *5 et n(1–p) = 10* $\geq$ *5*

*alors la fréquence empirique* $F_n$ *suit approximativement une loi normale*

$$F_n \underset{approx}{\sim} \mathcal{N}(0{,}75;0{,}0685)$$

➢ *quantile d'ordre 95% de la fréquence*
$0{,}75 + z_{0,95} \times 0{,}0685 =$
$0{,}75 + 1{,}645 \times 0{,}0685 =$
$0{,}75 + 0{,}113 = 0{,}863$

➔ *95% des échantillons de* **X** *issus de* $\mathcal{P}$ *de taille* ***n*** = *40 ont une fréquence de réussite inférieure à 86,3%*

④ *une fréquence de réussite supérieure à 86,3% observée sur un échantillon de taille* ***n*** = *40 sera considéré comme "étonnamment" élevé*

# Exemple variable qualitative (2)

- **Exemple : résultats au DEUG en 1994**

*échantillon de* ***X*** *issu de* $\mathcal{P}$ *de taille*

***n*** *= 40 (****n*** $\geq$ *30)*

*np = 30* $\geq$ *5 et n(1–p) = 10* $\geq$ *5*

*alors la fréquence empirique* $F_n$ *suit approximativement une loi normale*

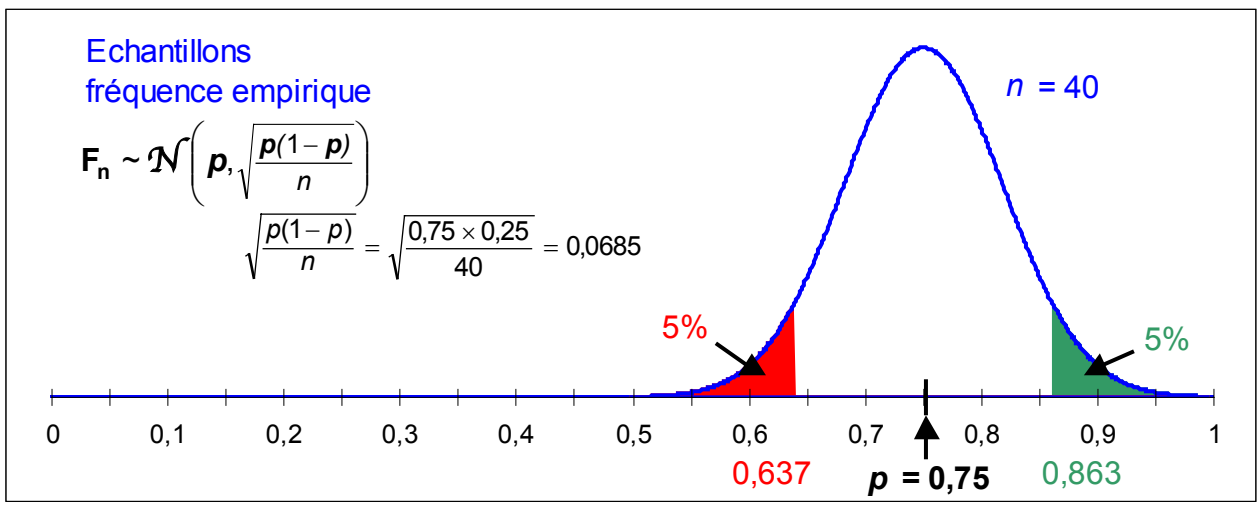
$$F_n \underset{approx}{\sim} \mathcal{N}(0{,}75;0{,}0685)$$

➢ *quantile d'ordre 5% de la fréquence*

$0{,}75 - z_{0{,}95} \times 0{,}0685 =$
$0{,}75 - 0{,}113 = 0{,}637$

➔ *5% des échantillons de* **X** *issus de* $\mathcal{P}$ *de taille* ***n*** *= 40 ont une fréquence de réussite inférieure à 63,7%*

④ *une fréquence de réussite inférieure à 63,7% observée sur un échantillon de taille* ***n*** *= 40 sera considéré comme "étonnamment" faible*



# 3.7 Intervalles de variation de la fréquence empirique (1)

**X variable qualitative dichotomique de $\mathcal{P}$ dans E = { oui, non }**

$\Rightarrow$un paramètre ***connu*** dans $\mathcal{P}$

**proportion de *"oui"* = *p***

échantillon de **X** issu de $\mathcal{P}$ de taille ***n***

- rechercher un intervalle dans lequel on a une chance $(1-\alpha)$ de trouver l'estimation de la proportion ***p*** sur un échantillon de taille ***n***

$\Rightarrow$cet intervalle est l'**intervalle de variation** (ou **intervalle de pari**, ou **intervalle de fluctuation**) de la fréquence empirique $F_n$ à $(1-\alpha)$ ou au risque $\alpha$ noté $\mathbf{I_{1-\alpha}(F_n)}$

# 3.7 Intervalles de variation de la fréquence empirique (2)

- **Exemple : résultats au DEUG en 1994**

$\mathcal{P}$ = *{ étudiants inscrits en 2d année de DEUG S.H. à Nanterre en 1994}*

***X*** = *"réussite au DEUG" : oui, non*
*variable qualitative dichotomique*
*un paramètre* ***connu*** *dans* $\mathcal{P}$
***p*** = *proportion de réussite =* ***0,75***

*échantillon de* ***X*** *issu de* $\mathcal{P}$ *de taille* ***n*** = *40*

➢ *la proportion de réussite* ***p*** *est estimée par la fréquence de réussite observée* ***f***

⑤ *dans quel intervalle s'attend-on raisonnablement à trouver* ***f*** *la fréquence de réussite observée sur l'échantillon de taille* ***n*** = *40 ?*

➢ *si* "**raisonnablement**" *veut dire "dans 90% des cas" on cherchera l'intervalle de variation à 90% (au risque 10%) de la fréquence empirique* $F_n$

➢ *si* "**raisonnablement**" *veut dire "dans 95% des cas" on cherchera l'intervalle de variation à 95% (au risque 5%) de la fréquence empirique* $F_n$

# 3.7 Intervalles de variation de la fréquence empirique (3)

X variable **qualitative dichotomique** de $\mathcal{P}$
dans E = { oui, non }
de **proportion de *"oui"*** = $p$ dans $\mathcal{P}$

échantillon de **X** issu de $\mathcal{P}$ de taille $n \geq 30$
avec $np \geq 5$ et $n(1-p) \geq 5$

- la fréquence empirique $F_n$ suit approximativement un modèle normal de moyenne $p$ et d'écart-type $\sqrt{p(1-p)/n}$

$$F_n \underset{approx}{\sim} \mathcal{N}\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

$\Rightarrow$l'intervalle de variation à $(1-\alpha)$ ou au risque $\alpha$ de $F_n$ s'écrit :

$$I_{1-\alpha}(F_n) \approx \left[p \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}}\right]$$

où $z_{1-\frac{\alpha}{2}}$ quantile d'ordre $1-\frac{\alpha}{2}$ de $Z \sim \mathcal{N}(0,1)$

# Exemple variable qualitative (1)

- **Exemple : résultats au DEUG en 1994**

*échantillon de **X** issu de $\mathcal{P}$ de taille*

***n** = 40 (**n** ≥ 30)*

*np = 30 ≥ 5 et n(1–p) = 10 ≥ 5*

*alors la fréquence empirique $F_n$ suit approximativement une loi normale*

$$F_n \underset{approx}{\sim} \mathcal{N}(0{,}75;0{,}0685)$$

➢ *intervalle de variation à 90% (au risque 10%) de la fréquence de réussite :*

$$I_{90\%}(F_n) \approx \left[ p \pm z_{0,95} \sqrt{\frac{p(1-p)}{n}} \right]$$

$$\approx \left[ 0{,}75 \pm 1{,}645 \sqrt{\frac{0{,}75 \times 0{,}25}{40}} \right]$$

$$\approx [0{,}75 \pm 0{,}113] = [0{,}637\,;\,0{,}863]$$

➔ *90% des échantillons de **X** issus de $\mathcal{P}$ de taille **n** = 40 ont une fréquence de réussite comprise entre 63,7% et 86,3%*

⑤ *dans 90% des cas, sur un échantillon de taille **n** = 40 on s'attend à observer une fréquence de réussite comprise entre 63,7% et 86,3%*

# Exemple variable qualitative (2)

- **Exemple : résultats au DEUG en 1994**

*échantillon de **X** issu de $\mathcal{P}$ de taille*

***n** = 40 (**n** ≥ 30)*

*np = 30 ≥ 5 et n(1–p) = 10 ≥ 5*

*alors la fréquence empirique $F_n$ suit approximativement une loi normale*

$$F_n \underset{approx}{\sim} \mathcal{N}(0{,}75;0{,}0685)$$

➢ *intervalle de variation à 95% (au risque 5%) de la fréquence de réussite :*

$$I_{95\%}(F_n) \approx \left[ p \pm z_{0,975} \sqrt{\frac{p(1-p)}{n}} \right]$$

$$\approx \left[ 0{,}75 \pm 1{,}96 \sqrt{\frac{0{,}75 \times 0{,}25}{40}} \right]$$

$$\approx [0{,}75 \pm 0{,}134] = [0{,}616 ; 0{,}884]$$

➔ *95% des échantillons de **X** issus de $\mathcal{P}$ de taille **n** = 40 ont une fréquence de réussite comprise entre 61,6% et 88,4%*

⑤ *dans 95% des cas, sur un échantillon de taille **n** = 40 on s'attend à observer une fréquence de réussite comprise entre 61,6% et 88,4%*

# Exemple variable qualitative (3)

- **Exemple : résultats au DEUG en 1994**

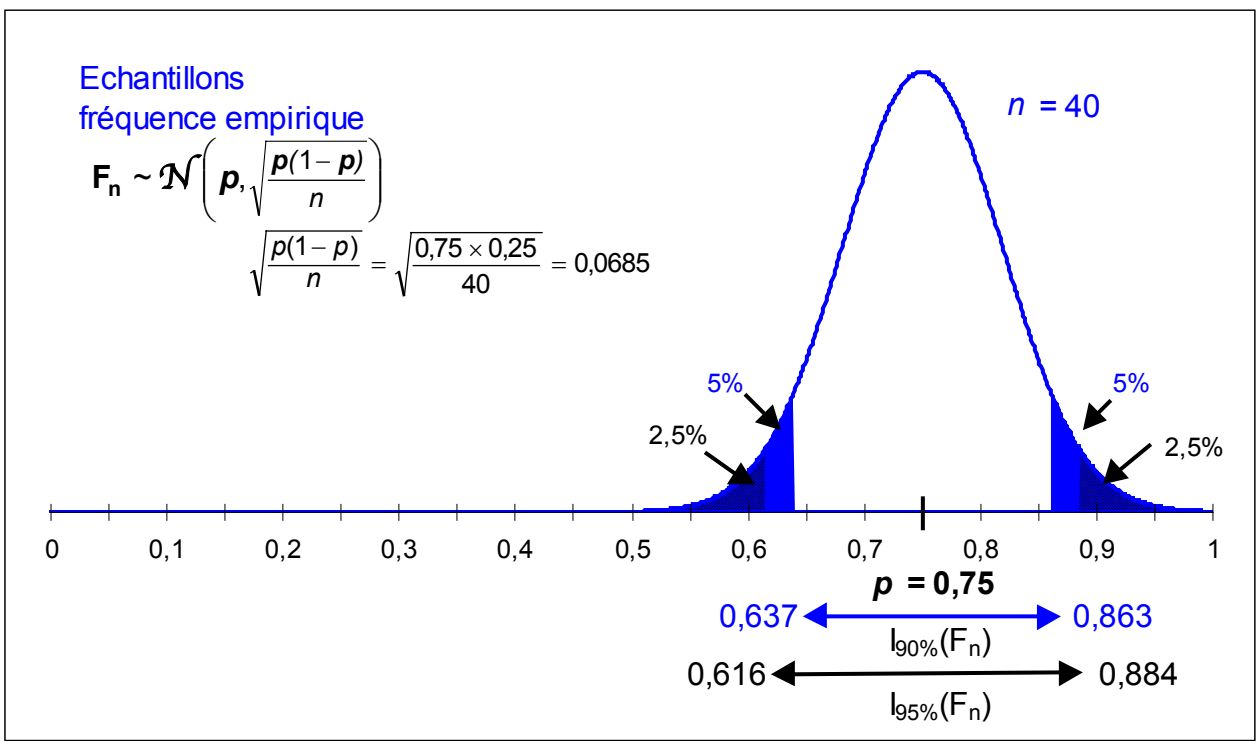
*échantillon de* ***X*** *issu de* $\mathcal{P}$ *de taille*

***n*** *= 40*

*la fréquence empirique* $F_n$ *suit approximativement une loi normale*

$$F_n \underset{approx}{\sim} \mathcal{N}(0{,}75;0{,}0685)$$

- *intervalle de variation à 90% (au risque 10%) de la fréquence de réussite :* $I_{90\%}(F_n) \approx [0{,}637\,;\,0{,}863]$
- *intervalle de variation à 95% (au risque 5%) de la fréquence de réussite :* $I_{95\%}(F_n) \approx [0{,}616\,;\,0{,}884]$



# Exemple variable qualitative (4)

- **Exemple : résultats au DEUG en 1994**

$\mathcal{P}$ = { *étudiants inscrits en 2d année de DEUG S.H. à Nanterre en 1994*}

***X*** = *"réussite au DEUG" : oui, non*
*variable qualitative dichotomique*

*échantillon de* ***X*** *issu de* $\mathcal{P}$ *de taille*

***n*** = 30 $np = 22,5 \geq 5$ et $n(1-p) = 7,5 \geq 5$

$$F_n \underset{approx}{\sim} \mathcal{N}(0,75;0,0791)$$

➢ *intervalle de variation à 95% (au risque 5%) de la fréquence de réussite :*

$$I_{95\%}(F_n) \approx \left[0,75 \pm 1,96\sqrt{\frac{0,75 \times 0,25}{30}}\right]$$

$$\approx [0,75 \pm 0,155] = [0,595\,;\,0,905]$$

***n*** =100 $np = 75 \geq 5$ et $n(1-p) = 25 \geq 5$

$$F_n \underset{approx}{\sim} \mathcal{N}(0,75;0,0433)$$

➢ *intervalle de variation à 95% (au risque 5%) de la fréquence de réussite :*

$$I_{95\%}(F_n) \approx \left[0,75 \pm 1,96\sqrt{\frac{0,75 \times 0,25}{100}}\right]$$

$$\approx [0,75 \pm 0,085] = [0,665\,;\,0,835]$$

# Exemple variable qualitative (5)

- **Exemple : résultats au DEUG en 1994**

*échantillon de* ***X*** *issu de* $\mathcal{P}$ *de taille*

***n*** *= 30* $F_n \underset{approx}{\sim} \mathcal{N}(0{,}75;0{,}0791)$

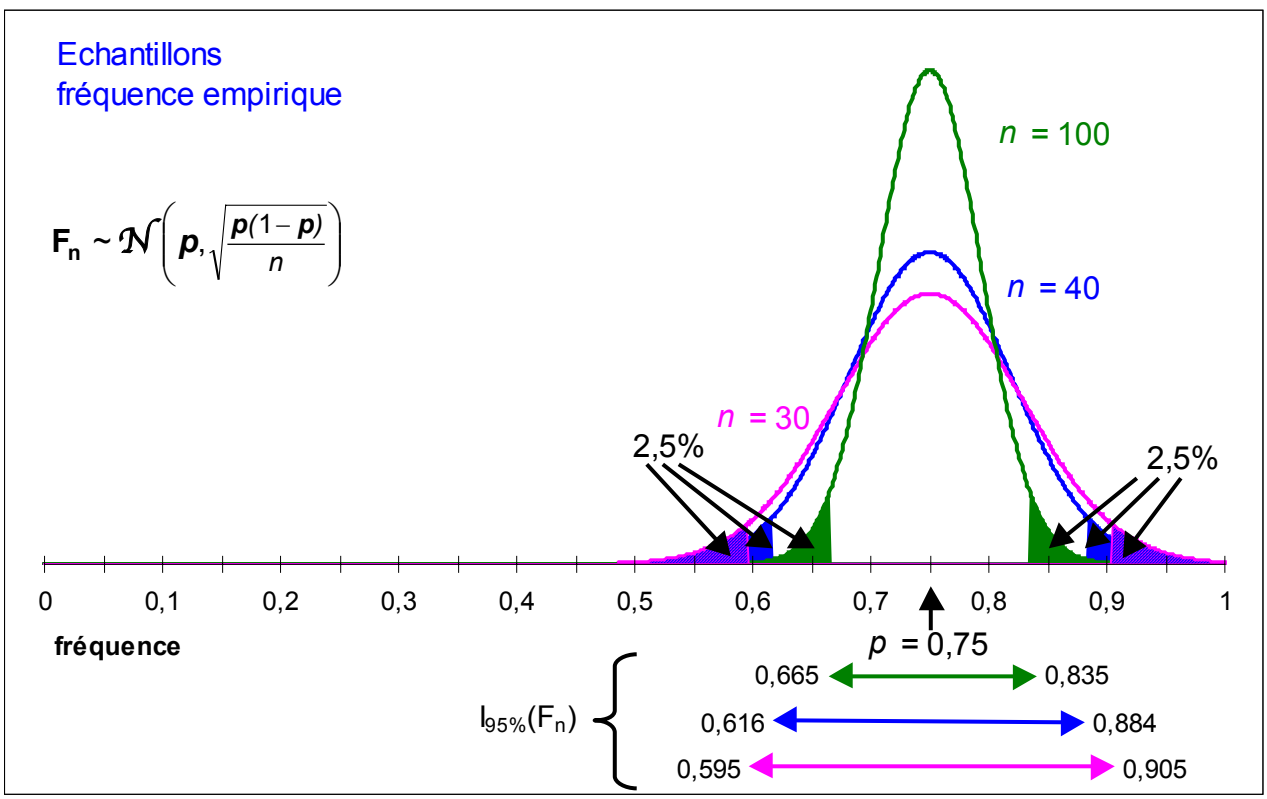
$I_{95\%}(F_n) \approx [0{,}75 \pm 0{,}155] = [0{,}595\ ;\ 0{,}905]$

***n*** *= 40* $F_n \underset{approx}{\sim} \mathcal{N}(0{,}75;0{,}0685)$

$I_{95\%}(F_n) \approx [0{,}75 \pm 0{,}134] = [0{,}616\ ;\ 0{,}884]$

***n*** *=100* $F_n \underset{approx}{\sim} \mathcal{N}(0{,}75;0{,}0433)$

$I_{95\%}(F_n) \approx [0{,}75 \pm 0{,}085] = [0{,}665\ ;\ 0{,}835]$



# 3.7 Intervalles de variation de la fréquence empirique (4)

## Remarques

- tous les intervalles de variation de la fréquence empirique sont centrés sur la proportion $p$ dans la population $\mathcal{P}$
- l'intervalle de variation dépend
  - **de la valeur du paramètre *p***
  - **du niveau $(1-\alpha)$ ou du risque $\alpha$**
  - **de la taille de l'échantillon *n***
- l'intervalle de variation ne dépend pas
  - **de la fréquence observée *f***
- l'intervalle de variation à $(1-\alpha)$ peut s'écrire :

$$I_{1-\alpha}(F_n) \approx \left[ p \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}} \right] = [p \pm e]$$

$e$ est la **demi-longueur** de l'intervalle $2e$ est l'amplitude ou longueur de l'intervalle

# Interprétation d'un intervalle de variation (1)

- un intervalle de variation est *fixe* : ses bornes sont fixes, calculées une fois pour toute, pour tous les échantillons de taille $n$
- une fréquence observée $f$ est *variable* : elle change pour chaque échantillon
- l'intervalle de variation au niveau $(1-\alpha)$ (au risque $\alpha$) est déterminé pour que :
  - $(1-\alpha)\%$ des valeurs de $F_n$ (des fréquences observées $f$) appartiennent à l'intervalle

  et
  - $\alpha\%$ des valeurs de $F_n$ (des fréquences observées $f$) n'appartiennent pas à l'intervalle

## Interprétation d'un intervalle de variation (2)

- l'intervalle de variation au niveau 95% (au risque 5%) est déterminé pour que :
  - **95% des valeurs de $F_n$ (des fréquences observées *f*) appartiennent à l'intervalle**

  **et**

  - **5% des valeurs de $F_n$ (des fréquences observées *f*) n'appartiennent pas à l'intervalle**
- par construction, par exemple sur **100** échantillons de taille ***n*** de **X** issus de $\mathcal{P}$, en moyenne :
  - **95 échantillons auront une fréquence observée *f* qui appartiendra à l'intervalle de variation à 95%**

  **et**

  - **5 échantillons auront une fréquence observée *f* qui n'appartiendra pas à l'intervalle de variation à 95%**

# Interprétation d'un intervalle de variation (3)

- **Exemple : résultats au DEUG en 1994**

  $\mathcal{P}$ = { *étudiants inscrits en 2d année de DEUG S.H. à Nanterre en 1994*}

  ***X*** = *"réussite au DEUG" : oui, non*
  *variable qualitative dichotomique*

  *un paramètre* ***connu*** *dans* $\mathcal{P}$ :
  ***p*** = *proportion de réussite* = ***0,75***

  *échantillon de* ***X*** *issu de* $\mathcal{P}$ *de taille* ***n*** = *40 : on observe* ***f*** = *88%*

  $I_{95\%}(F_n) \approx [0{,}75 \pm 0{,}134] = [0{,}616\,;\,0{,}884]$

  *donc* ***f*** *appartient à l'intervalle de variation à 95%*

  ① *échantillon d'étudiants* tiré au sort dans $\mathcal{P}$ *: on a observé* un *parmi les 95% des échantillons qui ont une fréquence de réussite observée dans l'intervalle de variation à 95%*

  ② *échantillon d'étudiants* non tiré au sort dans $\mathcal{P}$ *: on ne peut pas mettre en doute le fait qu'il est représentatif des étudiants de 2d année de DEUG S.H. à Nanterre en 1994 pour la réussite au DEUG*

# Interprétation d'un intervalle de variation (4)

- **Exemple : résultats au DEUG en 1994**

  $\mathcal{P}$ = *{ étudiants inscrits en 2d année de DEUG S.H. à Nanterre en 1994}*

  ***X*** = *"réussite au DEUG" : oui, non variable qualitative dichotomique*
  ***p*** = *proportion de réussite =* ***0,75***

  *échantillon de* ***X*** *issu de* $\mathcal{P}$ *de taille* ***n*** = *40 : on observe* ***f*** = *60%*

  $I_{95\%}(F_n) \approx [0{,}75 \pm 0{,}134] = [0{,}616\,;\,0{,}884]$

  *donc* ***f*** *n'appartient pas à l'intervalle de variation à 95%*

  ① *échantillon d'étudiants* tiré au sort dans $\mathcal{P}$ *: on a observé* un *parmi les 5% des échantillons pour lesquels f est en dehors de* $I_{95\%}(F_n)$

  ② *étudiants d'un groupe de TD* (non tiré au sort dans $\mathcal{P}$ ) *: a priori non représentatif et les observations le confirment*

  ③ *échantillon d'étudiants tiré au sort en 2005 :* $\mathcal{P}$ *= { étudiants en 2005 } la proportion de réussite au DEUG n'est plus la même*

# 3.8 Précision dans l'estimation de la proportion

La demi-longueur ***e*** de l'intervalle de variation à $(1-\alpha)$ représente la **précision** à $(1-\alpha)$ dans l'estimation de la proportion ***p*** sur un échantillon de taille ***n***, ou la **marge d'erreur** d'échantillonnage au risque $\alpha$

$$\boldsymbol{e} = z_{1-\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{\boldsymbol{n}}}$$

où $z_{1-\frac{\alpha}{2}}$ quantile d'ordre $1-\frac{\alpha}{2}$ de $\mathbf{Z} \sim \mathcal{N}(0,1)$

- la précision ***e*** dépend de ***p***
- pour ***n*** fixé, la précision ***e*** ne change pas lorsque l'échantillon change
- pour ***n*** fixé, plus le niveau $(1-\alpha)$ augmente (le risque $\alpha$ diminue) plus la marge d'erreur augmente ou la précision diminue
- plus la taille de l'échantillon ***n*** augmente, plus la marge d'erreur diminue ou la précision augmente

# Exemple variable qualitative (1)

- **Exemple : résultats au DEUG en 1994**

$\mathcal{P}$ = { *étudiants inscrits en 2d année de DEUG S.H. à Nanterre en 1994*}

***X*** = *"réussite au DEUG" : oui, non*
*variable qualitative dichotomique*
*un paramètre* ***connu*** *dans* $\mathcal{P}$
***p*** = *proportion de réussite =* ***0,75***

*échantillon de* ***X*** *issu de* $\mathcal{P}$ *de taille*
***n*** = *40 (****n*** $\geq$ *30)*
*np = 30* $\geq$ *5 et n(1–p) = 10* $\geq$ *5*

*alors la fréquence empirique* $F_n$ *suit approximativement une loi normale*

$$F_n \underset{approx}{\sim} \mathcal{N}(0{,}75;0{,}0685)$$

➢ *intervalle de variation à 90% (au risque 10%) de la fréquence de réussite :*

$$I_{90\%}(F_n) \approx \left[0{,}75 \pm 1{,}645\sqrt{\frac{0{,}75 \times 0{,}25}{40}}\right]$$

$$\approx [0{,}75 \pm 0{,}113]$$

➔ *la marge d'erreur à 90% dans l'estimation de la proportion de réussite sur les échantillons de taille* ***n*** = *40 est de 11,3%*

# Exemple variable qualitative (2)

- **Exemple : résultats au DEUG en 1994**

*échantillon de* ***X*** *issu de* $\mathcal{P}$ *de taille*

***n*** *= 40 (****n*** $\geq$ *30)*

*np = 30* $\geq$ *5 et n(1–p) = 10* $\geq$ *5*

*alors la fréquence empirique* $F_n$ *suit approximativement une loi normale*

$$F_n \underset{approx}{\sim} \mathcal{N}(0{,}75;0{,}0685)$$

➢ *intervalle de variation à 95% (au risque 5%) de la fréquence de réussite :*

$$I_{95\%}(F_n) \approx \left[0{,}75 \pm 1{,}96\sqrt{\frac{0{,}75 \times 0{,}25}{40}}\right]$$

$$\approx [0{,}75 \pm 0{,}134]$$

➔ *la marge d'erreur à 95% dans l'estimation de la proportion de réussite sur les échantillons de* **X** *issus de* $\mathcal{P}$ *de taille* ***n*** *= 40 est de 13,4%*

# Exemple variable qualitative (3)

- **Exemple : résultats au DEUG en 1994**

$\mathcal{P}$ = { *étudiants inscrits en 2d année de DEUG S.H. à Nanterre en 1994*}

***X*** = *"réussite au DEUG" : oui, non*
*variable qualitative dichotomique*

*échantillon de* ***X*** *issu de* $\mathcal{P}$ *de taille*

$n = 30$ ($n \geq 30$)
$np = 22{,}5 \geq 5$ et $n(1-p) = 7{,}5 \geq 5$

*alors la fréquence empirique* $F_n$ *suit approximativement une loi normale*

$$F_n \underset{approx}{\sim} \mathcal{N}(0{,}75;0{,}0791)$$

➢ *intervalle de variation à 95% (au risque 5%) de la fréquence de réussite :*

$$I_{95\%}(F_n) \approx \left[0{,}75 \pm 1{,}96\sqrt{\frac{0{,}75 \times 0{,}25}{30}}\right]$$

$$\approx [0{,}75 \pm 0{,}155] = [0{,}595\,;\,0{,}905]$$

➔ *la marge d'erreur à 95% dans l'estimation de la proportion de réussite sur les échantillons de* **X** *issus de* $\mathcal{P}$ *de taille* $n = 30$ *est de 15,5%*

# Exemple variable qualitative (4)

- **Exemple : résultats au DEUG en 1994**

  $\mathcal{P}$ = { *étudiants inscrits en 2d année de DEUG S.H. à Nanterre en 1994*}

  ***X*** = *"réussite au DEUG" : oui, non*
  *variable qualitative dichotomique*

  *échantillon de* ***X*** *issu de* $\mathcal{P}$ *de taille*
  ***n*** *= 100* (***n*** $\geq$ *30)*
  $np = 75 \geq 5$ *et* $n(1-p) = 25 \geq 5$

  *alors la fréquence empirique* $F_n$ *suit approximativement une loi normale*

  $$F_n \underset{approx}{\sim} \mathcal{N}(0{,}75;0{,}0433)$$

  ➢ *intervalle de variation à 95% (au risque 5%) de la fréquence de réussite :*

  $$I_{95\%}(F_n) \approx \left[0{,}75 \pm 1{,}96\sqrt{\frac{0{,}75 \times 0{,}25}{100}}\right]$$

  $$\approx [0{,}75 \pm 0{,}085] = [0{,}665\,;\,0{,}835]$$

  ➔ *la marge d'erreur à 95% dans l'estimation de la proportion de réussite sur les échantillons de* **X** *issus de* $\mathcal{P}$ *de taille* ***n*** *= 100 est de 8,5%*

# Exemple variable qualitative (5)

- **Exemple : résultats au DEUG en 1994**

  *échantillon de* ***X*** *issu de* $\mathcal{P}$ *de taille*

  ***n = 30***

  $$I_{95\%}(F_n) \approx [0{,}75 \pm 0{,}155] = [0{,}595\,;\,0{,}905]$$

  ➢ *la marge d'erreur à 95% est de 15,5%*

  ***n = 40***

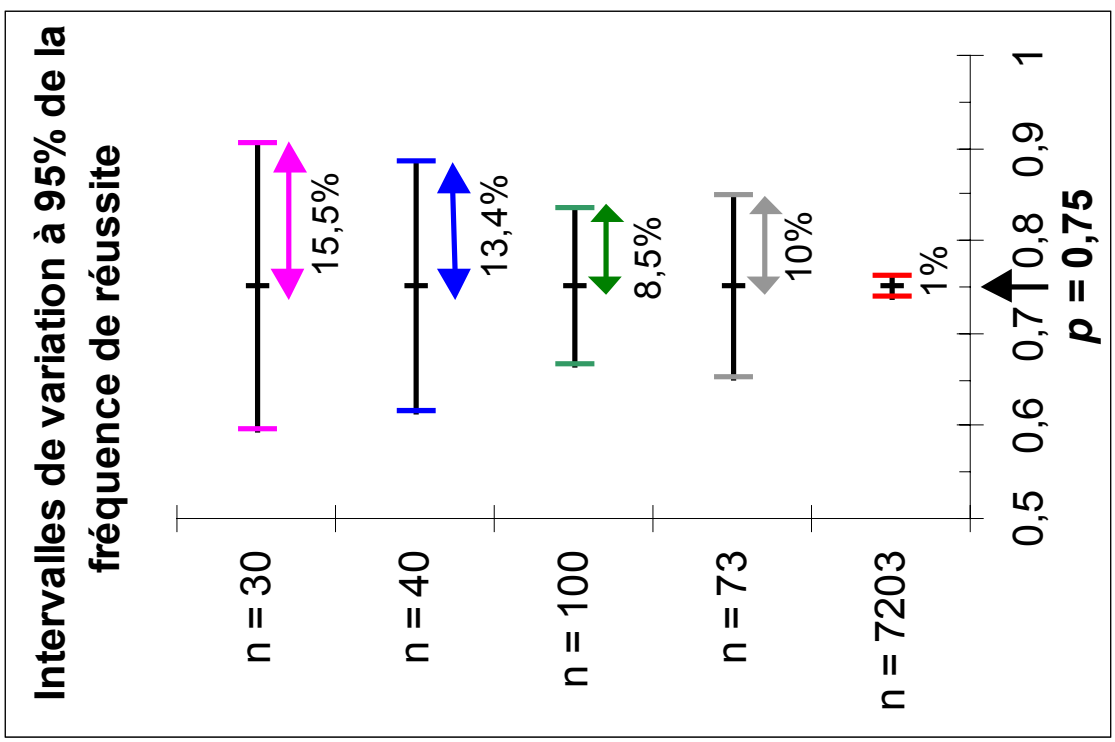
  $$I_{95\%}(F_n) \approx [0{,}75 \pm 0{,}134] = [0{,}616\,;\,0{,}884]$$

  ➢ *la marge d'erreur à 95% de 13,4%*

  ***n = 100***

  $$I_{95\%}(F_n) \approx [0{,}75 \pm 0{,}085] = [0{,}665\,;\,0{,}835]$$

  ➢ *la marge d'erreur à 95% de 8,5%*



# 3.9 Taille minimum de l'échantillon pour une précision minimum (1)

- Pour avoir une marge d'erreur à $(1-\alpha)$ inférieure à $e$ sur l'estimation de la proportion $p$ par la fréquence empirique :

$$z_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}} \leq e$$

la longueur (amplitude) de l'intervalle de variation au niveau $(1-\alpha)$ ou au risque $\alpha$ n'excède pas $2e$

⇒**la taille de l'échantillon $n$ vérifie :**

$$n \geq p(1-p)\left(\frac{z_{1-\frac{\alpha}{2}}}{e}\right)^2$$

➔pour obtenir une marge d'erreur d'au plus $e$ (ou une précision d'au moins $e$) il faut une taille d'échantillon d'au moins $n$

# 3.9 Taille minimum de l'échantillon pour une précision minimum (2)

## Remarques

- la taille minimum $n$ dépend de
  - **la proportion $p$ de la variable X**
  - **la marge d'erreur maximum $e$**
  - **le niveau $(1-\alpha)$ ou risque $\alpha$**
- la taille minimum $n$ augmente lorsque
  - **le produit $p(1-p)$ augmente**
  - **le niveau $(1-\alpha)$ augmente**
    **le risque $\alpha$ diminue**
  - **la marge d'erreur maximum $e$ diminue**
    **la précision minimum $e$ augmente**
- vérifier que pour la taille $n$ obtenue

$$n \geq 30 \qquad np \geq 5 \qquad n(1-p) \geq 5$$

# Exemple variable qualitative (1)

- **Exemple : résultats au DEUG en 1994**

  $\mathcal{P}$ = { *étudiants inscrits en 2d année de DEUG S.H. à Nanterre en 1994*}

  ***X*** = *"réussite au DEUG" : oui, non*
  *variable qualitative dichotomique*

  ***p*** = *proportion de réussite* = ***0,75***

  *échantillon de* ***X*** *issu de* $\mathcal{P}$ *de taille* ***n***

  *si* $\boldsymbol{n} \geq 30$ $\boldsymbol{n}p = 0{,}75 \times \boldsymbol{n} \geq 5$
  *et* $\boldsymbol{n}(1-p) = 0{,}25 \times \boldsymbol{n} \geq 5$

  *alors* $F_n \underset{approx}{\sim} \mathcal{N}\left(0{,}75; \sqrt{\frac{0{,}75 \times 0{,}25}{\boldsymbol{n}}}\right)$

# Exemple variable qualitative (2)

- **Exemple : résultats au DEUG en 1994**

➢ *pour* ***n*** *= 40 l'intervalle de variation à 90% (au risque 10%) de la fréquence de réussite :*

$$I_{90\%}(F_n) \approx [0{,}75 \pm 0{,}113]$$

➔ *la marge d'erreur à 90% dans l'estimation de la proportion de réussite sur les échantillons de taille* ***n*** *= 40 est de 11,3%*

➢ *marge d'erreur à 90% d'au plus 10%*

$$I_{90\%}(F_n) \approx \left[0{,}75 \pm 1{,}645\sqrt{\frac{0{,}75 \times 0{,}25}{n}}\right]$$

$$\approx [0{,}75 \pm 0{,}1]$$

$$1{,}645\sqrt{\frac{0{,}75 \times 0{,}25}{n}} \leq 0{,}1$$

$$\boldsymbol{n} \geq 0{,}75 \times 0{,}25\left(\frac{1{,}645}{0{,}1}\right)^2 = 50{,}74$$

***n*** *= 51 ≥ 30*
***np*** *= 0,75×51 = 38,25 ≥ 5*
***n***(1–*p*) *= 0,25×51 = 12,75 ≥ 5*

➔ *pour que la marge d'erreur à 90% dans l'estimation de la proportion de réussite n'excède pas 10% il faudra un échantillon d'**au moins** 51 étudiants*

# Exemple variable qualitative (3)

- **Exemple : résultats au DEUG en 1994**

➢ *pour* ***n*** *= 40 l'intervalle de variation à 95% (au risque 5%) de la fréquence de réussite :*

$$I_{95\%}(F_n) \approx [0{,}75 \pm 0{,}134]$$

➔ *la marge d'erreur à 95% dans l'estimation de la proportion de réussite sur les échantillons de taille* ***n*** *= 40 est de 13,4%*

➢ *marge d'erreur à 95% d'au plus 10%*

$$I_{95\%}(F_n) \approx \left[0{,}75 \pm 1{,}96\sqrt{\frac{0{,}75 \times 0{,}25}{n}}\right]$$

$$\approx [0{,}75 \pm 0{,}1]$$

$$n \geq 0{,}75 \times 0{,}25\left(\frac{1{,}96}{0{,}1}\right)^2 = 72{,}03$$

$n = 73 \geq 30$
$np = 0{,}75 \times 73 = 54{,}75 \geq 5$
$n(1-p) = 0{,}25 \times 73 = 18{,}25 \geq 5$

➔ *pour que la marge d'erreur à 95% dans l'estimation de la proportion de réussite n'excède pas 10% il faudra un échantillon d'***au moins** *73 étudiants*

# Exemple variable qualitative (4)

- **Exemple : résultats au DEUG en 1994**

➢ *marge d'erreur à 95% d'au plus 1%*

$$I_{95\%}(F_n) \approx \left[0{,}75 \pm 1{,}96\sqrt{\frac{0{,}75 \times 0{,}25}{n}}\right]$$

$$\approx [0{,}75 \pm 0{,}01]$$

$$\boldsymbol{n} \geq 0{,}75 \times 0{,}25\left(\frac{1{,}96}{0{,}01}\right)^2 = 7203$$

➔ *pour que la marge d'erreur à 95% dans l'estimation de la proportion de réussite n'excède pas 1% il faudra un échantillon d'**au moins** 7203 étudiants*