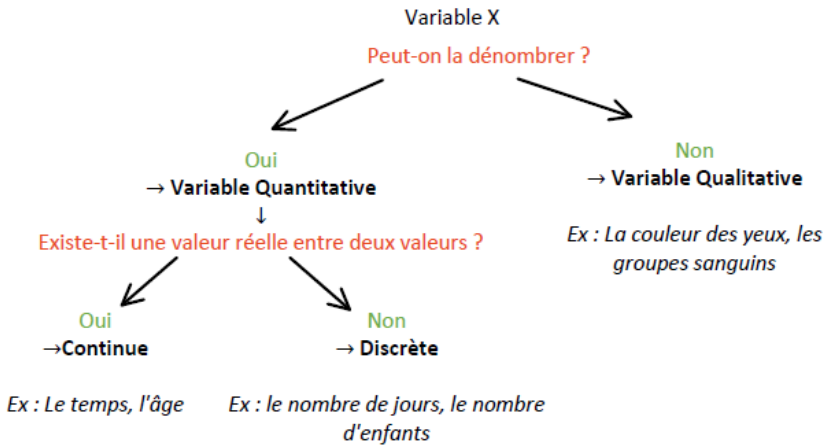
Fiche révisions TSR – UE 4 – Introduction aux Statistiques descriptives

Ces fiches ont été rédigées par des étudiants des années supérieures. Elles ne peuvent pas servir de support à une éventuelle contestation lors du concours de P.A.C.E.S.

N.B : Ce cours reprend des notions déjà connues de vos précédentes années (lycée, collège …). Il est donc l'un des plus simples, mais reste très important à connaître, puisqu'il sert de bases à la plupart des exercices de l'UE 4.

I. La Variable étudiée



Pour une Variable Quantitative Continue :

- La **représentation graphique** est un **Histogramme**
- On utilise les **densités de fréquence** calculées par : $D = \frac{Fréquence}{Largeur\ de\ Classe}$

Pour une Variable Quantitative Discrète :

- La **représentation graphique** est un **Diagramme à bâtons,** avec un polygône de fréquence
- On utilise les **fréquences**

II. Les Indicateurs de positions

Ils concernent **une valeur**

- **Mode :** C'est la valeur la plus fréquente de la distribution (elle peut être unique, on il peut y en avoir plusieurs)
- **Minimum**
- **Maximum**
- **Les Quartiles :** Q_1 sépare les 25% inférieurs des données. De la même manière, Q_3 sépare les 75% inférieurs des données. Pour les calculer/trouver, on utilise les Fréquences Cumulées Croissantes (F.C.C)
  - Dans le cas d'une variable quantitative **Discrète :**
    - Si on a une valeur qui « tombe pile », pas de problème !
    - Si les 25%, 75% … « tombent » entre deux valeurs, on additionne ces deux valeurs et on divise le tout par 2.
  - Dans le cas d'une variable quantitative **Continue,** on utilise la méthode de calcul suivante (donnée pour Q_3) :

$$Q_3 = Valeur\ inférieur\ de\ la\ classe + \frac{75 - Valeur\ inférieure\ de\ la\ classe}{Densité\ de\ fréquence\ de\ la\ classe}$$

N.B : - On procède de la même manière pour tous les calculs de Quartiles, percentiles, médianes, en changeant le **75 par la valeur souhaitée** (50 pour la médiane par exemple)

- On trouve des équivalences : par exemple, Q_2 est aussi la médiane, ou P_{50}.

- **Médiane :** Si on utilise les fréquences cumulées croissantes, elle partage l'échantillon en 2. Elle n'est **pas sensible aux valeurs extrêmes**
- **Moyenne :** On la définit m tel que : $m = \Sigma \frac{xi}{n}$ ou $m = \frac{\Sigma ni \times xi}{n}$ avec x_i apparait n_i fois, et $n = \Sigma ni$

III. Indicateurs de dispersion

Ils représentent un **écart entre les valeurs**

- **Etendue :** Maximum-Minimum
- **FCC :** probabilité d'individus $\leq$ Valeur donnée
  N.B : La bonne représentation de la FCC est la **marche d'escalier**
  Il existe de la même manière la FCD (décroissante)
- **Intervalle interquartile :** $Q_3 - Q_1$
- **Intervalle semi-interquartile :** $\frac{Q3 - Q1}{2}$

- **Variance & Ecart-Type :** On note S pour l'échantillon, et σ pour la population. S & σ représentent les écarts-type, et S^2 & σ^2 représentent les variances. Ainsi $S = \sqrt{S^2}$ et $\sigma = \sqrt{\sigma^2}$. Attention, pour **l'échantillon :**
  - **S_n^2 & S_n** concernent **l'échantillon exact :** ils sont **biaisés**
  - **S_{n-1}^2 & S_{n-1}** concernent **une estimation de l'échantillon :** c'est une correction de **S_n^2 & S_n**

  On note aussi que σ et σ^2 concernant la population, ils sont inaccessibles par calcul.
  Les calculs (les -1 sont à effectuer pour les calculs de S_{n-1}^2 & S_{n-1})

$$S(n-1)^2 = \frac{\Sigma\,(xi-m)^2}{n-1} \rightarrow S(n-1) = \sqrt{\frac{\Sigma\,(xi-m)^2}{n-1}}$$

Ou

$$S(n-1)^2 = \frac{1}{n-1}\left[\Sigma xi^2 - \frac{(\Sigma xi)^2}{n-1}\right]$$

Ou

$$S(n-1)^2 = \frac{1}{n-1}\left[\Sigma(\mathrm{ni} \times xi^2) - \frac{(\Sigma \mathrm{ni} \times xi)^2}{n-1}\right]$$

N. B : On estime que $\boldsymbol{Etendue \approx 4 \times \sigma^2}$

IV. Changement de variable linéaire & indicateur

Attention : il faut que ce soit la **même unité** !

- Ajout ? (X + b)
  - Indicateurs de **position décalés (de +b)**
  - Indicateurs de **dispersion** restent les **mêmes** !
- Multiplication ? (aX)
  - Indicateurs de **position multipliés par le facteur a**
  - Indicateurs de **dispersion multiplié par le facteur a**
- Changement linéaire ? (aX + b)
  - Indicateurs de **position** = $\boldsymbol{a \times (I.position) + b}$
  - Indicateurs de **dispersion** = $\boldsymbol{|a| \times (I.dispersion)}$

    Ex : moyenne : $m_y = am_x + b$

    Ecart-type : $S_y = |a|S_x$

    Variance : $S_y^2 = a^2S_x^2$

Fiche révisions TSR – UE 4 – Introduction aux probabilités

Ces fiches ont été rédigées par des étudiants des années supérieures. Elles ne peuvent pas servir de support à une éventuelle contestation lors du concours de P.A.C.E.S.

NB : Ce cours est pour l'essentiel un rappel de votre programme de Terminale S en probabilité. Il est destiné à vous « remettre dans le bain », ces bases sont des pré-requis nécessaires à la réalisation des exercices et s'intègrent à un programme plus vaste de probabilités en PACES.

I. Probabilité d'un événement

Ω correspond à l'univers : c'est l'ensemble des résultats possibles.

Un **événement E** correspond à une sous-partie d'**Ω.**

Le **cardinal**, noté « card », est peut-être une notion nouvelle pour certains d'entre vous. C'est tout simplement le **nombre d'issus qui réalise un événement.**

Ex : Dans un jeu de 52 cartes, si E correspond à « tirer un roi », alors card (E) = 4 (car on peut tirer le roi de cœur, le roi de pic, le roi de trèfle et le roi de carreau).
!! De ce fait, card (**Ω**) est le nombre total d'issus, dans cet exemple, card (Ω) = 52

De ces définitions, on déduit que **p (E) = card (E) / card (Ω)**

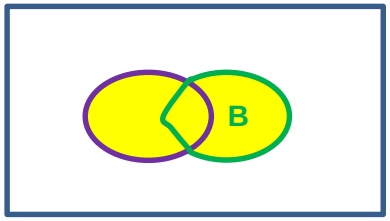
Ainsi :

- 0 < p (E) < 1 (car 1 = card (**Ω**) / card (**Ω**)),
- P (**Ω**) = 1

II. Quelques définitions

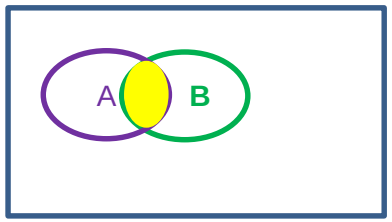
1) Union et intersection

L'union correspond aux issus qui vérifient un évènement A **OU** un événement B et se note (AUB).



Ainsi, (AUB) correspond à la partie jaune.

L'intersection correspond aux issus qui vérifient à la fois un évènement A ET un évènement B et se note (A∩B).



(A∩B) correspond à la partie colorée en jaune.

2) Compatibilité et incompatibilité

Deux **évènements incompatibles** ne peuvent se réaliser en même temps (or c'est la définition même de l'intersection). Ainsi, **p(A∩B) = 0** et **p (AUB) = p(A) + p (B).**
Ex : Sur un tirage, A = tirer un roi et B = tirer un as. A et B sont incompatibles.

Si deux évènements sont **compatibles, p (AUB) = P (A) + p (B) – p (A∩B)**
En effet, si l'on reprend notré schéma on ne peut se contenter d'additionner la probabilité de chacun des deux cercles car ils sont à cheval l'un sur l'autre, on n'oublie donc pas de soustraire la probabilité que les deux cercles se croisent.

3) Evènement contraire

$\bar{A}$ est l'évènement contraire de A. Ainsi, ils forment tous deux une partition de l'univers et **p(A) + p ($\bar{A}$) = 1**.

4) Probabilités conditionnelles

Ce sont les probabilités en **« sachant que ».** Par exemple, la probabilité de A sachant B se note p (A/B).

On a **p (A/B) = p (A∩B) / p (B).**

Dans la plupart des exercices, il est utile de dresser un tableau comme celui-ci.
(voir ci-dessous)

	A	$\overline{A}$
B	$A \cap B$	$\overline{A} \cap B$
$\overline{B}$	$\overline{B} \cap A$	$\overline{A} \cap \overline{B}$

On le remplit avec des **effectifs**.
Exemple :

	A	$\overline{A}$	N total
B	23	12	35
$\overline{B}$	8	16	24
N total	31	28	59

On a p (A/B) = 23/35.
Si on compare au tableau ci-dessus, on retrouve notre formule.
P $(A \cap B) = 23/59$
P (AUB) = (35+31-23) / 59

5) Indépendance

Si deux évènements sont indépendants, c'est que la réalisation de l'un n'influence pas sur la réalisation de l'autre. Ainsi, **p $(A \cap B)$ = p (A) x p (B).**
Cependant, on ne peut se fier à la logique pour juger de l'indépendance de deux évènements, il faut toujours vérifier par le calcul.

Par exemple, si la probabilité de l'évènement R : « porter un bonnet rouge » ne semble pas avoir d'influence sur V « porter une écharpe vert ». Si l'énoncé noud donne : p (R) = ¼ et p (V) = 1/ et que p $(V \cap R)$ = 1/16.
On a p (R) x p (V) = 1/12 ≠ 1/16 donc R et V ne sont pas indépendants.

Fiche révisions TSR – UE 4 – Evaluation d'un examen diagnostic

Ces fiches ont été rédigées par des étudiants des années supérieures. Elles ne peuvent pas servir de support à une éventuelle contestation lors du concours de P.A.C.E.S.

I. Sensibilité et spécificité

Ces valeurs s'appuient sur un test dichotomique avec test positif et négatifs. Dans presque tous vos exercices on vous présente la proportion M de malade et la proportion T de test positifs (avec $\bar{M}$ et $\bar{T}$ les témoins et les tests négatifs) avec :

	M (malades)	$\bar{M}$ (témoins)
T	Vrai positif	Faux positif
$\bar{T}$	Faux négatif	Vrai négatif

La **sensibilité** (Se) représente la capacité du test à détecter des maladies : elle est représentée par la proportion de tests positifs dans la population des malades.
Elle correspond à : $Se = P(T/M) = P(M \cap T) / P(M) = P(VP) / P(M)$.
$1 - Se = P(\bar{T}/M) = P(M \cap \bar{T}) / P(M) = P(FN) / P(M)$.

La **spécificité** (Sp) représente la capacité du test à ne pas représenter autre chose : elle est représentée par la proportion de test négatif dans la population de témoins. Elle correspond à : $Sp = P(\bar{T}/\bar{M}) = P(\bar{M} \cap \bar{T}) / P(\bar{M}) = P(VN) / P(\bar{M})$.
$1 - Sp = P(T/\bar{M}) = P(\bar{M} \cap T) / P(\bar{M}) = P(FP) / P(\bar{M})$.

Se et Sp sont par définition des valeurs comprises entre 0 et 1, elles doivent au minimum avoir une valeur supérieure à 0.5 pour être pertinente (plus qu'un lancer de pièce où on aura 50% de positifs et négatifs que l'on soit malade ou non).

Si **Se = 1**, il n'existe pas de faux négatif, un test négatif élimine obligatoirement la maladie, il est très utile en cas de dépistage.
Si **Sp = 1**, alors on n'a pas de faux positifs et un test positif confirme la maladie : on appelle ceci un **test pathognomonique** et est très utile dans le cadre d'un diagnostic.

On peut étudier la qualité globale du diagnostic à partir de 2 autres valeurs :

- L'**indice de Youden** : $IY = Se + Sp - 1$; il sert de comparaison à la pièce de monnaie où l'indice est de 0 (Se = Sp = 0.5).
- Le **rapport de vraisemblance** : $VM = \frac{Se}{1-Sp} = \frac{P(VP)}{P(FP)}$; il permet de réagir vis-à-vis d'un test positif, si le VM est de 5, quand on a un test positif il y a 5 fois plus de chances que le sujet soit malade que non malade.

II. Valeurs prédictives

La **valeur prédictive positive** représente la validité d'un test positif, elle représente la proportion de malades parmi les tests positifs. $VPP = P(M/T) = P(VP) / P(T)$.
On peut la déterminer à partir de Se, Sp et f (fréquence de la maladie) avec le théorème de Bayes : on peut voir que la VPP est **proportionnelle à Se, Sp et f**.

La **valeur prédictive négative** représente la validité d'un test négatif, elle représente la proportion de témoins parmi les tests négatifs. $VPN = P(\bar{M}/\bar{T}) = P(VN) / P(\bar{T})$.
On peut la déterminer à partir de Se, Sp et f avec le théorème de Bayes : on peut voir que VPN est **proportionnelle à Se et Sp** mais est **inversement proportionnelle à f**.

Attention : Se et Sp sont intrinsèques au test et seront invariables quel que soit le contexte. VPP et VPN tiennent compte de la fréquence de la maladie et donc du contexte.

Si on est face à un cas continu (hypertension artérielle) on le ramène à un cas dichotomique en posant un seuil de maladie (± 140 mmHg).
Quand on est positif au-dessus du seuil :

- Si le **seuil augmente**, on a moins de test positif donc les faux positifs diminuent → Sp augmente ; les faux négatifs augmentent → Se diminue.
- Si le **seuil diminue**, on a plus de test positifs donc les faux positifs augmentent → Sp diminue ; les faux négatifs diminuent → Se augmente.

Conseil n°1 : Si vous n'arrivez pas à vous souvenir de quelle probabilité correspond à quel terme, vous pouvez retrouver tout cela dans le tableau ci-dessous :

	M	$\bar{M}$	Total
T	Vrai positif	Faux positif	Test positif
$\bar{T}$	Faux négatif	Vrai négatif	Test négatif
Total	Malade	Témoin	1

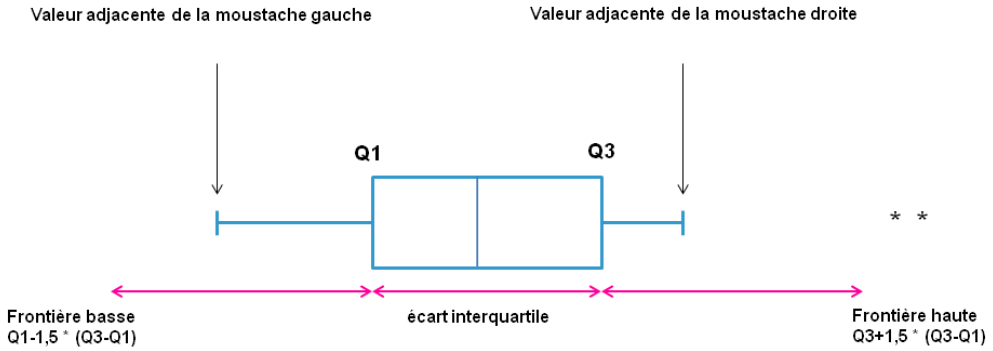
$Se = VP / M$
$Sp = VN / \bar{M}$
$1 - Se = FN / M$
$1 - Sp = FP / \bar{M}$
$VPP = VP / T$
$VPN = VN / \bar{T}$

Conseil n°2 : On aura beau vous faire plein d'exercices différents sur des tests en tout genre, au final ça reste pareil. Si vous n'y arrivez toujours pas, essayez de le rapprocher à un exercice du même genre que vous avez parfaitement compris comme un exercice de colle ou un ED que vous avez fait.

Fiche révisions TSR – UE 4 – Boîte à moustaches & Astuces des statistiques descriptives

Ces fiches ont été rédigées par des étudiants des années supérieures. Elles ne peuvent pas servir de support à une éventuelle contestation lors du concours de P.A.C.E.S.

I. Boîte à moustaches



Les **valeurs adjacentes correspondantes à :**

- **$1{,}5\ IQR + Q_3$** pour la valeur de **droite**
- **$Q_1 - 1{,}5\ IQR$** pour la valeur de **gauche**

Sur la **boîte,** on retrouve **:**

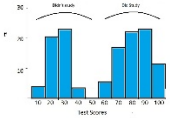
- La **médiane**
- Les **limites de la boîte** (rectangle) sont les **Q_1 et Q_3**

On parle de **valeurs atypiques** pour les valeurs qui sont **en dehors des valeurs adjacentes**.

Remarque : Si la **variable** répond à une distribution **gaussienne**, la **boîte** est **centrée par la médiane**.

II. Astuces

Ces astuces correspondent au premier cours.



La courbe est dite **bimodale** que si **deux groupes de valeurs qui les plus fréquentes** sont bien séparés dans les distributions

Pour une **variable quantitative continue**, pensez bien aux **densités de fréquences** notamment pour trouver le **mode** ! Si les **classes ont toutes la même largeur**, vous pouvez directement **chercher la classe modale sans passez par les densités de fréquences**.

Pour la **moyenne**, ne vous embêtez pas à la calculer à la main ! Elle est **automatiquement donnée** par les **calculatrices** (voir cours sur les calculatrices).

Variance et écart-type permettent de voir la **dispersion des données par rapport à la moyenne**. Attention à l'**unité de la variance** !! Elle est **au carré** !

Pour différencier S_n et S_{n-1} : **S_n** sous-estime la variance, il est donc normalement **plus petit que S_{n-1}.**

Pour savoir si la **distribution est normale**, on calcule **l'intervalle [m – 2s ; m + 2s]** et on regarde si **95% des valeurs** sont **dedans**.

Fiche révisions TSR – UE 4 – Fluctuation d'échantillonnage d'une proportion

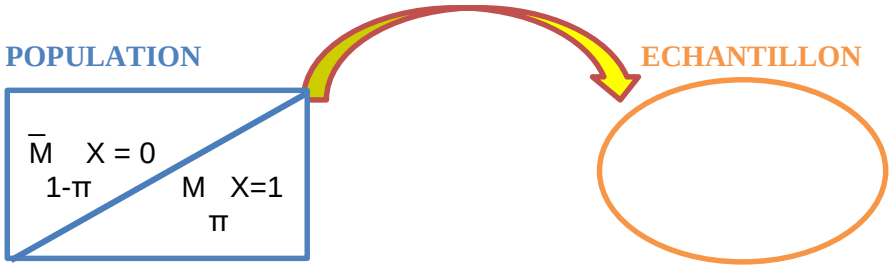
Ces fiches ont été rédigées par des étudiants des années supérieures. Elles ne peuvent pas servir de support à une éventuelle contestation lors du concours de P.A.C.E.S.

Dans ce cours, on part de la **population** pour arriver à **l'échantillon**.

Rappel de notation :
p : proportion observée dans un échantillon (issu de la population)
n : nombre de personnes dans l'échantillon (issu de la populaiton)
π : proportion réel du caractère étudié dans la population

Dans notre cas, pi est connu et on cherche p.

Soit **M** l'évènement : le sujet est malade. On associe à M une variable **X** qui peut prendre deux valeurs, car le sujet est soit malade, soit il ne l'est pas. X est donc une variable bde Bernouilli. Pour un résultat positif, mettons « le sujet est malade », on attribue à X la valeur 1, et dans le cas contraire, X=0.
Soit **N** la variable qui comptabilise le nombre de personnes malades dans l'échantillon. N correspond donc à la somme des X dans l'échantillon. Ainsi, N suit une loi Binomiale (répétition d'une variable de Bernoulli)



Rappelez vous, pour calculer une proportion, on a p = card (M) / n, c'est-à-dire le nombre de personne malade sur le nombre de personnes au total. Si on désigne P par la variable probabilité d'être malade dans l'échantillon, on a :

P = N/n

On a donc **E (P) = π et Var(P) = π (1- π) / n**

A CERTAINES CONDITIONS !!!!! , il est possible d'apparenter cette loi binomiale (quantitative discrète) s'apparente à une **loi normale** (quantitative continue).

Pour cela, on doit absolument vérifier les conditions suivantes :

$$\pi\, n \geq 5 \text{ et } n\,(1-\pi) \geq 5$$

Si et seulement si ces conditions sont vérifiées, on va pouvoir utiliser la loi normale pour un construire un **intervalle de fluctuation** IF (ou intervalle de probabilité IP) autour de **π** pour obtenir un encadrement de la proportion observée dans l'échantillon issue de la population.

Cet inter valle s'écrit :

$$IF(P) = \pi \pm z\alpha \sqrt{\pi\,(1-\pi)/n}$$

Z alpha dépend du risque α que l'on accepte.
Si l'on accepte α = 5%, on a 1- α = 1 -5% =95 % de chance que la proportion du caractère dans l'échantillon se trouve dans notre intervalle. On cherche le z α correspondant dans le tableau sur a loi normale qui est mis à votre disposition. Dans la majorité des cas, on nous donne α = 5% et z α = 1.96 ou α=1% et z α = 2,576. Je vous conseille de retenir ces deux valeurs pour gagne du temps.

A noter :

- Plus n est grand, plus p s'approche de pi
- Vérifiez toujours vos conditions avant de calculer l'intervalle
- On peut vous donner un intervalle, et vous demander de trouver le risque. Pour prendre un exemple simple :

Pi = 0,5 et IF = [0,4 ; 0,6] et n = 100

Ainsi : z alpha $\sqrt{0,5\,(1-0,5)/100}$ = 0,1
Donc z alpha = 2 et alpha = 5%

Fiche révisions TSR – UE 4 – Estimation d'une proportion π

Ces fiches ont été rédigées par des étudiants des années supérieures. Elles ne peuvent pas servir de support à une éventuelle contestation lors du concours de P.A.C.E.S.

Dans ce cours, on part de **l'échantillon** pour arriver à la **population**.

Ainsi, on connait **p** (qui est une réalisation de P : proportion du caractère dans une population), et on veut en déduire **π**.
Pour cela, il y a deux façons de procéder :

- **Estimation ponctuelle**
- **Estimation par intervalle de confiance**

I. Estimation ponctuelle

II est estimé par p car p fluctue autour de π c'est à dire que :

$$E(p) = \pi \text{ et } Var(p) = \pi(1-\pi)/n$$

Ainsi, on peut dire que p est une **bonne estimation** et une **estimation non biaisée** de **π.**

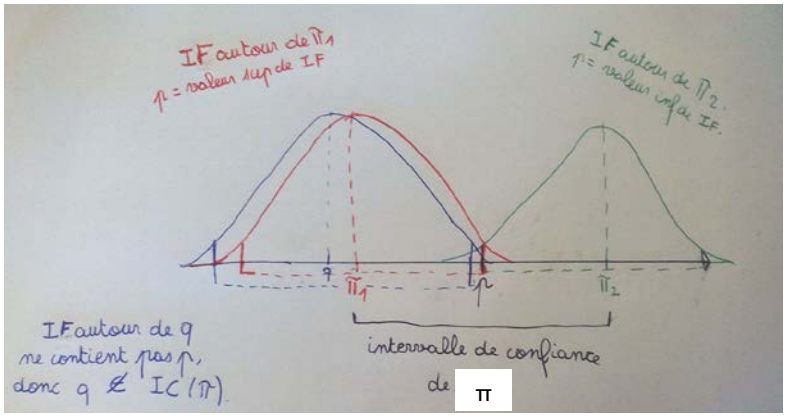
ATTENTION !! Si p=0,63, il est incorrect d'écrire π = 0,63. En effet π ≠ 0,63 pour la simple et bonne raison qu'on ne connait pas pi, on en fait juste une estimation. On écrit donc : **π est estimé par p=0,63.**
Ceci peut être un éventuel piège de QCM.

II. Estimation par intervalle de confiance.

On cherche à construire un **intervalle de confiance** de π autour de p noté **IC (π)** (différent de l'intervalle de fluctuation de p qui lui est construit autour de π).
Pour se faire, on recherche toutes les valeurs de π compatibles avec p. Ainsi :

- On choisit une valeur de π
- On construit un intervalle de fluctuation IF au risque voulu autour de cette valeur
- On vérifie si p appartient à cet intervalle
  - ➔ Si c'est le cas : pi est compatible avec p et cette valeur de pi appartient à IC
  - ➔ Sinon : la valeur choisie n'appartient pas à IC.

Ceci n'est pas à retenir mais vous aide à comprendre comment est construit un intervalle de confiance.



Dans la pratique, suite à différentes approximations, on a :

$$IC\ \alpha(\pi) = p \pm z\alpha \sqrt{p(1-p)/n}$$

Cependant, il faut toujours que certaines conditions soient validées pour utiliser cette loi, c'est-à-dire :

$$n\pi 1 \geq 5$$
$$n\pi 2 \geq 5$$
$$n(1-\pi 1) \geq 5$$
$$n(1-\pi 2) \geq 5$$

Cette fois, les conditions sont à vérifier a posteriori, avec les valeurs limites de l'intervalle trouvé.

ATTENTION !!! Il est faux de dire, pour un intervalle au risque α = 5%, que « il y a 95% de chance que π soit dans l'IC », on dit qu'il y a 95% de chance que IC contienne π » car c'est IC qui est variable et π qui est une valeur exacte, l'erreur ne vient pas de π mais de l'intervalle qu'on a construit.

Si les conditions ne sont pas vérifiées, il est possible d'utiliser la **table de la loi binomiale** pour construire notre intervalle de confiance.

Comment utiliser la table de la loi binomiale ?

Table -3-

INTERVALLE de CONFIANCE d'une PROPORTION

(au risque α de 5 %)

	PROPORTION OBSERVÉE p										
Effectif de l'échantillon	0 %	5 %	10 %	15 %	20 %	25 %	30 %	35 %	40 %	45 %	50 %
10	0 - 31		0 - 45		3 - 56		7 - 65		12 - 74		19 - 81
20	0 - 17	0 - 25	1 - 32	3 - 38	6 - 44	9 - 49	12 - 54	15 - 59	19 - 64	23 - 68	27 - 73
30	0 - 12		2 - 27		8 - 39		15 - 49		23 - 59		31 - 69
40	0 - 9	1 - 17	3 - 24	6 - 30	9 - 36	13 - 41	17 - 47	21 - 52	25 - 57	29 - 62	34 - 66
50	0 - 7		3 - 22		10 - 34		18 - 45		26 - 55		36 - 64
60	0 - 6	1 - 14	4 - 21	7 - 27	11 - 32	15 - 38	19 - 43	23 - 48	28 - 53	32 - 58	37 - 63
70	0 - 5		4 - 20		11 - 31		20 - 42		28 - 52		38 - 62
80	0 - 5	1 - 12	4 - 19	8 - 25	12 - 30	16 - 36	20 - 41	25 - 46	29 - 52	34 - [illegible]	39 - 61
90	0 - 4		5 - 18		12 - 30		21 - 41		30 - 51		[illegible] - 61
100	0 - 4	2 - 11	5 - 18	9 - 24	13 - 29	17 - 35	21 - 40	26 - 45	30 - 50	35 - 55	40 - 60
160	0 - 2	2 - 10	6 - 16	10 - 22	14 - 27	19 - 32	23 - 38	28 - 43	32 - 48	37 - [illegible]	[illegible] - 58
200	0 - 2	2 - 9	6 - 15	10 - 21	15 - 26	19 - 32	24 - 37	28 - 42	33 - 47	38 - [illegible]	[illegible] - 57

Deux points à noter :

- Les proportions indiquées sont comprises entre et %, au-delà de %, on travaille avec le complémentaire.
  Ex : n=80 et p=70%
  Pour p = 30%, pi est compris entre 20 et 41%, donc pour p=70%, pi est compris entre 59 et 80 %.
  Et IC = [0,59 ; 0,8]
- Pour une valeur d'échantillon qui n'est pas proposée, on fait une moyenne des intervalles des échantillons qui l'encadre.
  Ex : n=25 et p = 0,2

IC = [0,7 ; 0,415]

A noter que la table qui vous est donnée est valable pour un risque α = %.

Comment déterminer le nombre de sujet nécessaire pour une pyrécision voulue ?

$IC\ \alpha\ (\pi) = p \pm z\ \alpha \sqrt{p\ (1-p)\ /n}$

$= p \pm \mathbf{i}$

i correspond au **demi – intervalle**, il caractérise la **précision**, plus i est petit, plus l'intervalle est précis.

On a $i = z\ \alpha \sqrt{p\ (1-p)\ /n}$

D'où $n = z\ \alpha^2\ p\ (1-p)\ /\ i^2$

Cette formule n'est pas donnée dans le formulaire mais facile à retrouver par une simple isolation de l'inconnue.

On voit dans la formule que n est fonction de :

- i → Si i augmente, n diminue (c'est logique, pour un plus grand demi-intervalle, on a une plus petite précision et donc on n'a pas besoin de réaliser l'expérience sur autant de sujets)
- α → Si α diminue, z α augmente donc n augmente (encore une fois, pour un plus petit risque il faut plus de personnes car on veut plus de certitude)
- p : nécessité d'avoir une idée sur la valeur de fréquence à estimer.

I. Fluctuation d'échantillonnage d'une moyenne observée

Soit X une variable aléatoire quantitative discrète ou continue caractérisée par sa moyenne μ et sa variance σ^2.
On prend un échantillon et on note M la moyenne observée. Elle correspond aussi à une variable aléatoire :

- Si on tire un individu, on lit la VA : M = X/1 ; E (M) = μ et Var (M) = σ^2.
- Si on tire n individus, on observe n VA X_i indépendantes → on observe ainsi $M = \frac{\Sigma Xi}{n}$ correspondant à un VA représentant la moyenne de n VA indépendantes → E (M) = μ et Var (M) = $\frac{\sigma^2}{n}$.
- M fluctue donc autour de μ et de $\frac{\sigma^2}{n}$.

Loi de distribution de M :

- Si $n \geq 30$ ou $X \sim \mathcal{N}(\mu, \sigma^2)$ → $\mathbf{M \sim \mathcal{N}(\mu, \sigma^2/n)}$.
- On peut déterminer l'intervalle de fluctuation au risque α (IF_α) avec :

$$\mathbf{IF_\alpha\ (M) = \mu \pm z_\alpha \sqrt{\sigma^2/n}}$$

- P (M ∈ IF) = 1 – α.

Dans la majorité des cas on choisira des cas où α = 5% (z_α = 1.96) ou α = 1% (z_α = 2.576). Il est conseillé de retenir ces valeurs car ça peut faire gagner du temps (et on sait que c'est très important en stats).

Exemple : On prend une population de moyenne μ = 250, σ^2 = 140 et qu'on tire au hasard un échantillon de n = 50 personnes, on recherche l'intervalle de fluctuation de la moyenne m au risque 5% :

$IF_{5\%}(m) = 250 \pm 1.96 \sqrt{\frac{140}{50}} = 250 \pm 3.28 = [246.72 ; 253.28]$

II. Fluctuation d'échantillonnage d'une variance observée

Soit une population X comparable à une variable aléatoire et caractérisée par une moyenne μ et une variance σ^2. On choisit un échantillon avec s_n^2 la variance observée dans cet échantillon, correspondant elle aussi à une variable aléatoire :

- $\mathbf{E\ (s_n^2) = \frac{n-1}{n} \sigma^2}$ → fluctue en-dessous de σ^2.

On détermine ainsi $\mathbf{s_{n-1}^2 = S^2 = \frac{n}{n-1} S_n^2}$ où $\mathbf{E\ (s^2) = \sigma^2}$.

- On considère que s^2 est une **estimation** de la variance σ^2.
- Cela correspond à xσn-1 dans la calculatrice.
- Attention : $s_n^2 \neq s^2$ (piège très facile à poser en qcm).

Exemple : On prend un échantillon de n = 40 individus où l'on observe une variance de s_n^2 = 100 : $s^2 = \frac{40}{39} s_n^2 =$ 102.56

Conseil : Ce chapitre est assez simple mais il reste pourtant très présent au concours, il serait vraiment dommage de ne pas le maîtriser, une simple erreur de distinction entre s_n^2 et s^2 et c'est tout l'exercice qui tombe à l'eau.
L'intervalle de fluctuation a normalement été vu en Terminale donc il ne pose pas trop de problèmes, de plus la formule est présente dans le formulaire.

Fiche révisions TSR – UE 4 – Estimation d'une moyenne et d'une variance

Ces fiches ont été rédigées par des étudiants des années supérieures. Elles ne peuvent pas servir de support à une éventuelle contestation lors du concours de P.A.C.E.S.

Soit une population X avec une moyenne μ et une variance σ^2 inconnues. On tire un échantillon où l'on observe une moyenne m et une variance s_n^2.
Comment estimer μ et σ^2 à partir de m et s_n^2 ?

I. Estimation ponctuelle

La moyenne μ de la population est estimée à partir de m (moyenne observée = $\frac{\Sigma xi}{n}$), m est une estimation non biaisée → elle fluctue autour de μ avec une variance de $\frac{\sigma^2}{n}$).

Pour la variance, s_n^2 est une mauvaise estimation car elle fluctue en-dessous en σ^2 : elle est biaisée. On va préférer utiliser $s^2 = \frac{n}{n-1} s_n^2$ qui fluctue autour de σ^2 et qui est une estimation non biaisée de σ^2.

Exemple : On tire un échantillon de 150 individus avec m = 10 et $s_n^2 = 30$. On peut estimer μ par m : $\mu \approx 10$; par contre on estime σ^2 par $s^2 = \frac{150}{149} \times 30 = 30.2$: $\sigma^2 \approx 30.2$.

II. Estimation de μ par un intervalle de confiance

Le but ici est de déterminer toutes les valeurs de μ dans la population compatibles avec l'observation : si on fait un intervalle de fluctuation au risque α si $n \geq 30$ ou $X \sim N(\mu ; \sigma^2)$ on doit pouvoir vérifier que m appartient à cet intervalle.
On définit ainsi μ_1 et μ_2 tels que m soit égal aux bornes inférieures puis supérieures de If_α de μ_1 et μ_2 : $m = \mu_1 + z_\alpha \sqrt{\sigma^2/n} = \mu_2 - z_\alpha \sqrt{\sigma^2/n}$.

$$\mathbf{IC\alpha\ (\mu) = [\mu 1\ ;\ \mu 2] = m \pm z_\alpha \sqrt{\sigma^2/n}}$$

On établit cet intervalle de confiance si $n \geq 30$ ou $X \sim N(\mu ; \sigma^2)$.
Problème : one ne connait pas σ^2.

III. Estimation de σ^2 pour établir un intervalle de confiance

On considère une estimation de σ^2 par s^2.
Si $M \sim \mathcal{N}(\mu ; \frac{\sigma^2}{n}) \rightarrow \mathbf{Z = \frac{M-\mu}{\sqrt{\sigma^2/n}} \sim \mathcal{N}(0 ; 1)}$

On appelle une variable $T = \frac{M-\mu}{\sqrt{s^2/n}} \sim$ **Loi de Student à (n-1) degrés de liberté**.

La loi de Student est un équivalent de la loi normale dont la représentation est plus aplatie en fonction du degré de liberté : plus le degré est faible plus la courbe est aplatie et l'intervalle de confiance sera large pour avoir un risque α donné. Une loi normale correspond à une loi de Student à ∞ d.d.l. (degrés de liberté).

On utilise (n-1) degrés de liberté et pas n degrés de liberté car dans la définition d'une loi normale et d'une loi de Student : $\Sigma (x_i - m) = 0$ → la dernière valeur doit dépendre des autres pour que la somme soit nulle.

A partir du moment où $n \geq 30$ (29 d.d.l.) on considère que l'on a presque affaire à une loi normale où le $t_{n-1, \alpha}$ peut se confondre avec z_α.

On peut donc donner une estimation de μ à partir de s^2 per l'intervalle de confiance :

$$\mathbf{IC_\alpha\ (\mu) = m \pm t_{n-1, \alpha} \sqrt{s^2/n}}$$

IV. Taille d'un échantillon

La taille n de l'échantillon influence la précision i telle que $IC_\alpha\ (\mu) = m \pm i$.

$\mathbf{i = t_{n-1, \alpha} \sqrt{s^2/n}}$ → n dépend de $\mathbf{s^2}$; de $\mathbf{t_{n-1, \alpha}}$ (qu'on peut confondre avec z_α si $n \geq 30$) ; et n dépend aussi de **i** (précision souhaitée).

Au final on obtient : $\mathbf{n = \frac{z\alpha^2 \times s^2}{i^2}}$

Fiche révisions TSR – UE 4 - Théorie statistique de décision – Test d'hypothèse et de signification

Ces fiches ont été rédigées par des étudiants des années supérieures. Elles ne peuvent pas servir de support à une éventuelle contestation lors du concours de P.A.C.E.S.

Il s'agit là de prendre une décision concernant la population, à partir des infos de l'échantillon.

I. Les étapes

- Choix de l'hypothèse à vérifier : on l'appellera **H0**
- Choix de l'hypothèe rivale qui sera acceptée si la nulle est rejetée : **H1**
- Définition d'une rège pour prendre la décision d'accepter ou de rejeter H0 à **un risque α**
- **Calcul sur un échantillon** aléatoire de la statistique appropriée
- **Prise de décision**

ATTENTION !! H0 et H1 sont définies pour la **POPULATION**, ces paramètres ne peuvent être définis à partir de l'échantillon.

Exemples : H0 : p = 0,7 → H1 : p ≠ 0,7 : **BILATERAL** car p peut être supérieur OU inférieur à 0,7. C'est ce modèle bilatéral qu'on utilise en PACES.
A noter qu'une définition de H1 unilatérale pourrait correspondre à p > 0,7

Pour prendre une décision : on rejette H0 si les valeurs de l'échantillon sont **significativement différentes.**

II. Les erreurs

Il existe 2 types d'erreur :

- Erreur de première espèce : **α**
  α = p (rejeter Ho si Ho est vrai) : c'est l'équivalent en probabilité d'un faux négatif. Cela correspond au degrès de signification. Il est fixé par l'utilisateur.
- Erreur de seconde espèce **β**
  β = p (accepter Ho si Ho est fausse) = (accepter h0 si H1 est vraie) : c'est l'équivalent en probabilité d'un faux positif.
  Cela correspond à un **manque de puissance du test.**
  En effet, **1-β = puissance du test**

On note que α et β évoluent dans des sens opposés. Si α augmente, β diminue.
De plus, pour un risque α fixé, β diminue si n augmente.

	Ho acceptée (non rejetée)	H0 rejetée (H1 acceptée)
H0 vraie	PAS D'ERREUR P (accepter H0 si H0 est vraie) = **1-α**	FAUX NEGATIF P (rejeter h0 si H0 est vrai) **α**
H0 fausse	FAUX POSITIF P (accepter H0 si H0 est fausse) **β**	PAS D'ERREUR P (rejeter H0 si H0 est fausse) = **1-β**

ATTENTION !! On ne peut pas conclure que Ho est vraie car on ne connait pas β.
Cependant, on peut conclure que H0 est fausse car on a fixé nous même le risque α.

Fiche révisions TSR – UE 4 – Comparaison d'une proportion observée et d'une proportion théorique

Ces fiches ont été rédigées par des étudiants des années supérieures. Elles ne peuvent pas servir de support à une éventuelle contestation lors du concours de P.A.C.E.S.

Exemple : Effet secondaire d'un médicament : $\pi_0 = 6\%$ → on présente un nouveau produit ayant passé des tests cliniques sur un échantillon de 1000 personnes où l'on trouve 3 effets indésirables (p = 4.3%). On émet 2 hypothèses :

- H_0 : "Il n'y a aucune modification : $\pi = \pi_0$".
- H_1 : "Il y a une modification de la fréquence des effets secondaires : $\pi \neq \pi_0$".

Si H_0 est vraie ($\pi = 6\%$) et $n\pi = 60 \geq 5$; $n(1-\pi) = 940 \geq 5$:

- $\mathbf{P \sim \mathcal{N}\left(\pi\ ;\ {\pi(1-\pi)}/{n}\right)}$ et $\mathbf{Z = \frac{P-\pi}{\sqrt{{\pi(1-\pi)}/{n}}} \sim \mathcal{N}\ (0\ ;\ 1)}$.

I. Estimation à partir d'un intervalle de fluctuation à partir de p

On réalise un **intervalle de fluctuation autour de π** et on vérifie si le p observée appartient à cet intervalle ou non → on accepte H_0 ou non.

Rappel : $IF_\alpha\ (P) = \pi \pm \sqrt{{\pi(1-\pi)}/{n}}$.

Exemple : $IF_{5\%}\ (P) = 0.06 \pm 1.96 \sqrt{{0.06\ x\ 0.94}/{1000}} = 0.06 \pm 0.015 = [0.045\ ;\ 0.075]$ → $p = 0.043 \notin IF_{5\%}\ (P)$ → **H_0 est rejetée** au risque $\alpha = 5\%$.

II. Estimation à partir d'un intervalle de fluctuation à partir de Z

On réalise un **intervalle de fluctuation autour de 0** et on vérifie si la variable z appartient à l'intervalle $[-z_\alpha\ ;\ z_\alpha]$ → on accepte H_0 ou non → **on rejette H_0 si $z \geq z_\alpha$**.

Rappel : $z = \frac{P-\pi}{\sqrt{{\pi(1-\pi)}/{n}}}$.

Exemple : Ici on prend un risque α de 5% soit un intervalle $IF_{5\%}\ (Z) = [-1.96\ ;\ 1.96]$.
$z = \frac{|0.043-0.06|}{\sqrt{{0.06\ x\ 0.94}/{1000}}} = 2.26 > 1.96$ → **H_0 est rejetée** au risque α de 5%.

On peut même déterminer le degré de signification à partir de la table de la loi normale → $2.26 > 2.170 = z_{3\%}$ → H_0 est même rejetée au risque α de 3%.

III. Estimation à partir d'un intervalle de confiance

On réalise un **intervalle de confiance autour de la proportion p observée** et on vérifie si la proportion π théorique appartient à cet intervalle.
Attention : l'intervalle de confiance pour l'estimation de π est le seul test où **les conditions se vérifient a postériori**.

Rappel : $IC_\alpha\ (\pi) = p \pm z_\alpha \sqrt{{p(1-p)}/{n}}$.

Exemple : $IC_{5\%}\ (\pi) = 0.043 \pm 1.96 \sqrt{{0.043\ x\ 0.957}/{1000}} = 0.043 \pm 0.013 = [0.030\ ;\ 0.056]$
→ $\pi = 0.06 \notin IC_{5\%}\ (\pi)$ → **H_0 est rejetée** au risque $\alpha = 5\%$.
$0.03 \times 1000 = 30 \geq 5$ → les conditions sont vérifiées a postériori.

Attention : Si lors du test on ne rejette pas H_0 ($p \in IF_{5\%}\ (P)$; $z \leq 1.96$; $\pi \in IC_{5\%}\ (\pi)$) cela ne veut pas forcément dire que l'on accepte H_0 car on ne connait pas le risque β d'accepter H_0 si H_0 est faux → le test manque de puissance, on ne peut pas conclure. On pourrait peut-être finalement rejeter H_0 si on augmentait la taille de l'échantillon.

Conseil n°1 : Il faut faire attention à bien lire l'énoncé de la question pour bien choisir la méthode à utiliser car généralement on ne vous en demandera qu'une et chacune des méthodes peut donner des résultats différents.

Conseil n°2 : Faites très attention au sujet des **intervalles de confiance** quand on vous demande un écart à 5% car il existe 2 méthodes de le faire : soit on le calcule comme c'est montré ci-dessus soit on utilise le **tableau donné dans le formulaire** qui reste plus précis et on peut vous poser un cas limite où H_0 est accepté dans une méthode et pas dans l'autre et vous poser le piège.
Quand vous avez un doute : privilégiez le tableau du formulaire (surtout quand vous obtenez un p et un n très proches des valeurs données du tableau) mais attention car cela ne concerne que le risque α de 5%.

Fiche révisions TSR – UE 4 – Comparaison de deux proportions observées

Ces fiches ont été rédigées par des étudiants des années supérieures. Elles ne peuvent pas servir de support à une éventuelle contestation lors du concours de P.A.C.E.S.

Exemple : On cherche à comparer les effets de 2 médicaments lors de 2 essais cliniques différents :

- Le médicament A a eu un taux de succès (p_1) de 60% sur 200 malades.
- Le médicament B a eu un taux de succès (p_2) de 50% sur 150 malades.
- Le médicament A est-il vraiment plus efficace ?

On pose les hypothèses suivantes :

- H_0 : "Les 2 médicaments ont un taux de succès identique : $\pi_1 = \pi_2$".
- H_1 : "Les 2 médicaments ont un taux de succès différent : $\pi_1 \neq \pi_2$".

Pour cela on va déterminer 2 variables :

- $P_1 \sim \mathcal{N}\left(\pi_1 ; \frac{\pi1(1-\pi1)}{n1}\right)$ et $P_2 \sim \mathcal{N}\left(\pi_2 ; \frac{\pi2(1-\pi2)}{n2}\right)$

Si H_0 est vraie ; $P_1 = P_2$ et $\mathbf{P_1 - P_2 = 0}$.
Rappel : $\mathbf{E(X \pm Y) = E(X) \pm E(Y)} \rightarrow E(P1-P2) = \pi_1 - \pi_2$
$\mathbf{Var(X \pm Y) = Var(X) + Var(Y)}$ si X et Y sont indépendants.
$\rightarrow Var(P_1 - P_2) = \frac{\pi1(1-\pi1)}{n1} + \frac{\pi2(1-\pi2)}{n2}$.

Si H_0 est vraie $\rightarrow \pi_1 = \pi_2$ et $P_1 - P_2 \sim \mathcal{N}\left(0 ; \pi(1-\pi)\left(\frac{1}{n1} + \frac{1}{n2}\right)\right)$
$\rightarrow \mathbf{Z} = \frac{\mathbf{P1-P2-0}}{\sqrt{\boldsymbol{\pi(1-\pi)}\left(\frac{1}{n1}+\frac{1}{n2}\right)}} \sim \mathcal{N}(0 ; 1) \rightarrow$ **<u>Problème</u>** : on ne connait pas $\pi \rightarrow$ on va l'estimer par $\mathbf{p = \frac{n1p1+n2p2}{n1+n2}} \rightarrow$ on va donc comparer $\mathbf{z = \frac{|P1-P2|}{\sqrt{p(1-p)\left(\frac{1}{n1}+\frac{1}{n2}\right)}}}$ à z_α pour accepter H_0 ou pas. (H_0 est rejetée si $|z| \geq z_\alpha$ (1.96 si on prend un risque α de 5%).

Exemple : $p = \frac{200 \times 0.6 + 150 \times 0.5}{200+150} = 0.56 \rightarrow z = \frac{|0.6-0.5|}{\sqrt{0.56(1-0.56)\left(\frac{1}{200}+\frac{1}{150}\right)}} = 1.86 < 1.96$.

<u>Conclusion</u> : la proportion du succès du médicament A n'est pas significativement supérieure au médicament B.

Attention lors du jugement de signification voire de causalité, lors d'un essai thérapeutique il y des variations sur les sujets, les critères d'exclusion et d'inclusion, la clause d'équivalence (allergie au médicament ≠ échec du médicament). Lors du tirage au sort il faut établie des groupes "compatibles" → intérêt d'utiliser les mêmes modalités de traitement en double aveugle.

Exemple lors des enquête cas témoin su la consommation d'alcool en lien avec la cirrhose du foie : on observe une plus grande proportion de cirrhose chez les patients consommant de l'alcool mais on ne peut pas établie un lien de causalité car on ne sait pas si <u>l'alcool est la cause ou la conséquence de la maladie</u> → facteur non contrôlé : on a besoin de connaître leur consommation avant la maladie → alcool = facteur de risque.

<u>Conseils</u> : Il n'est pas nécessaire de connaitre toutes les formules par cœur, le formulaire est là pour ça mais le plus important est de comprendre comment on arrive à cette formule car ce sera beaucoup plus facile de l'utiliser après.

Cependant les petites formules comme l'espérance et la variance d'une somme est à connaître et vous sera probablement demandé au concours sans formulaire.

Le mieux est de se baser sur un exemple et de bien le comprendre pourquoi on fait ces calculs car les cas seront exactement les mêmes, il y a juste les chiffres qui vont changer. Le fait de rapprocher l'exercice que vous avez en face de vous à un exercice que vous connaissez et que vous maîtrisez peut vous aider à mieux le résoudre.

Fiche révisions TSR – UE 4 – Comparaison d'une moyenne observée à une moyenne théorique

Ces fiches ont été rédigées par des étudiants des années supérieures. Elles ne peuvent pas servir de support à une éventuelle contestation lors du concours de P.A.C.E.S.

I. Principe et Généralités

Dans quelle situation utilise-t-on cette comparaison ?

- ➔ Il faut **n observations** à partir de laquelle on peut calculer une **moyenne observée.**
- ➔ On doit aussi disposer d'une **moyenne théorique** (moyenne d'une population donnée)

Quelles sont les conditions ?

- ➔ On va considérer comme dans beaucoup de test que la variable étudiée suit une loi normale, ainsi on aura :
  - o **$X \sim N(\mu, \sigma^2)$**
  - o Ou **$n \geq 30$**

Correspond aussi à **$M \sim N(\mu, \frac{\sigma^2}{n})$**

Quelles hypothèses formuler ?

- ➔ On considère que **E(M) = μ_0**
- ➔ Alors **$H_0 : \mu = \mu_0$** et **$H_1 : \mu \neq \mu_0$** ou si l'hypothèse est **unilatéral : $H_1 : \mu >/< \mu_0$**

Attention : On parle des populations, pas de l'échantillon ! Il faut considérer qu'on parle d'une vérité qu'on va ensuite démontrer (on établit la vérité dans l'hypothèse et ensuite on la démontre par le calcul à l'aide des données de l'échantillon).

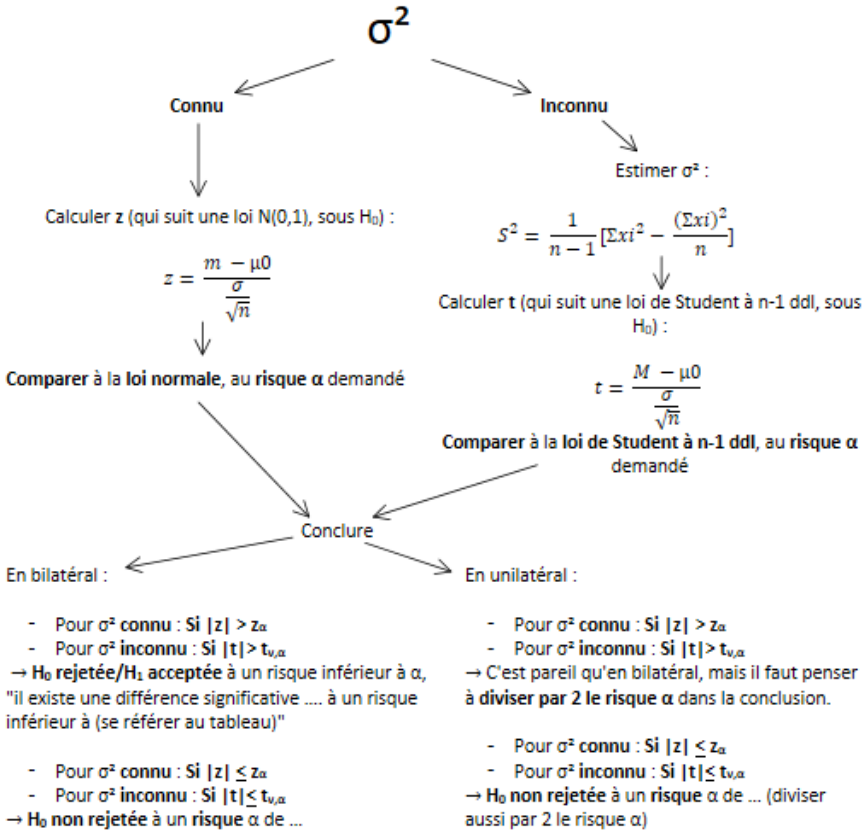
De façon générale, on calcule $Z = \frac{M-\mu 0}{\sqrt{\frac{\sigma^2}{n}}} = \frac{M-\mu 0}{\frac{\sigma^2}{\sqrt{n}}}$ $avec\ Z \sim N(0,1)$ **sous H_0**

- ➔ Ainsi, on considère que **$\mu_0 = \mu$**. On fait donc finalement la **différence** entre la **moyenne théorique** (traduite par cette égalité), et la **moyenne observée** (m)

II. Mise en application

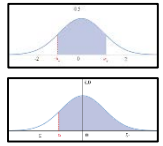
La **mise en application** dépend de la **connaissance ou non de σ^2**

Voici un petit schéma pour vous aider :



Explication de l'unilatéral :

- Si on décide de faire un test avec des hypothèses en unilatéral, le problème est que nous avons des tables qui sont en bilatéral pour les tests pratiqués.
- On va contourner le problème : en bilatéral, le risque est « reparti » à droite et à gauche de la courbe. Disons qu'on fait l'hypothèse en unilatéral de **$H_1 : \mu < \mu_0$.** Dans ce cas, la partie qui nous intéresse est à gauche de la courbe. On fait le test et on veut par exemple que le risque α **en unilatéral** soit de **5%.**
  - o Si on prend dans le tableau un risque α de 5% (en bilatéral), on aura 2,5% de chaque côté de la courbe.
  - o Il faut donc prendre un risque α de 10% en bilatéral pour avoir un risque de 5% de chaque côté de la courbe.



- A quel risque conclure ? Il faut préciser ! On peut dire « risque α de 5% en unilatéral » qui est plus souvent utilisé **quand on fait de l'unilatéral** que « risque α de 10% en bilatéral »

III. Estimation d'un IC

Précédemment, nous avons fait de l'estimation ponctuelle. Il est possible de calculer un intervalle de confiance à partir de la moyenne observée et de comparer cet intervalle avec la moyenne théorique.

En pratique :

- Les **conditions** : **X~N ou n ≥ 30**
- Les **lois** s'appliquent selon la **connaissance** ou non de **σ^2** :
  - Si **σ^2 connue** : Loi **Normale**
  - Si **σ^2 inconnue** : Loi de **Student**
- L'**estimation** de **σ^2** se fait par **S^2**

Les **calculs** sont les suivants :

- Si **σ^2 connue** : $IC = \left[m - z_\alpha \times \frac{\sigma}{\sqrt{n}} \; ; \; m + z_\alpha \times \frac{\sigma}{\sqrt{n}}\right]$
  - ➔ IC au risque α, basé sur la moyenne
  - ➔ Cet IC comprend (1- α)%
- Si **σ^2 inconnue** : $IC = [m - t_{\nu,n-1} \times \frac{s}{\sqrt{n}} \; ; \; m + t_{\nu,n-1} \times \frac{s}{\sqrt{n}}]$

IV. Quelques remarques …

- Si les **ddl** de la loi de Student **> 100**, on peut dire que la **loi se rapproche de la loi Normale**.
- Si n augmente, la probabilité de mettre en évidence une différence **si elle existe** est augmentée : on parle de **puissance augmentée** du test.
- **Intervalle de confiance** et **Estimation ponctuelle** partagent :
  - Les **conditions**
  - L'**estimation de σ^2 par S^2**
  - Les **lois s'appliquant selon la connaissance de σ^2** (Loi N si connue, Loi de Student si inconnue)

Fiche révisions TSR – UE 4 – Comparaison de deux moyennes observées

Ces fiches ont été rédigées par des étudiants des années supérieures. Elles ne peuvent pas servir de support à une éventuelle contestation lors du concours de P.A.C.E.S.

I. Principe & Généralités

Dans quelle situation utilise-t-on cette comparaison ?

- ➔ Il faut **deux échantillons.**
- ➔ On doit aussi avoir les **deux moyennes des deux échantillons.**

Quelles sont les conditions ?

- ➔ On va considérer comme dans beaucoup de test que les variables étudiées suivent chacune une loi normale, ainsi on aura :
  - o $X_1 \sim N(\mu_1, \sigma_1^2)$
  - o Ou $n_2 \geq 30$

  Correspond aussi à $M_1 \sim N_1\left(\mu_1, \frac{\sigma 1^2}{n1}\right)$
  - o $X_2 \sim N(\mu_2, \sigma_2^2)$
  - o Ou $n_2 \geq 30$

  Correspond aussi à $M_2 \sim N_2\left(\mu_2, \frac{\sigma 2^2}{n2}\right)$

Quelles hypothèses formuler ?

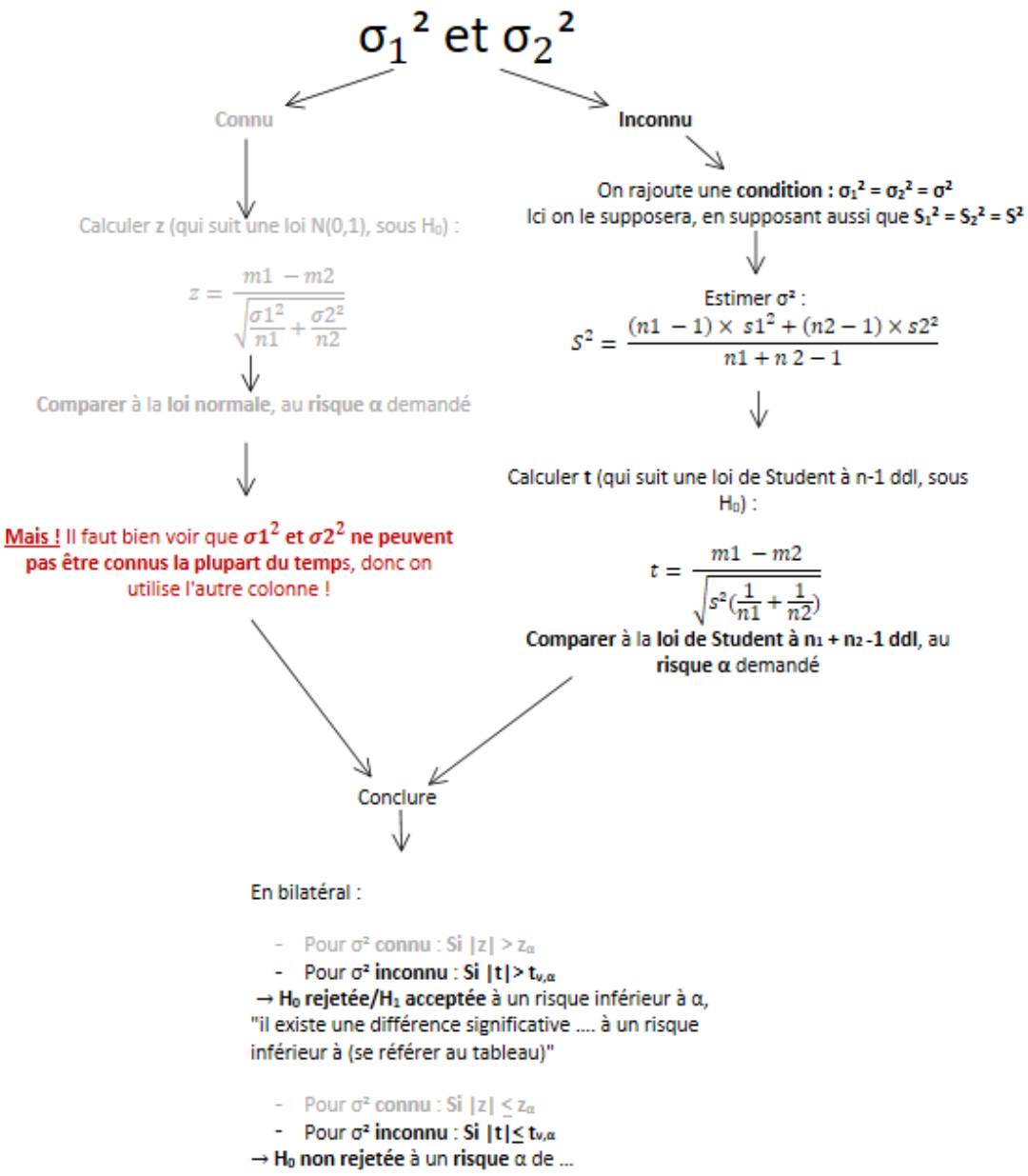
- ➔ $H_0 : \mu_1 = \mu_2$ et $H_1 : \mu_1 \neq \mu_2$
- ➔ *En fait, sous H_0, on considère que M_1 et M_2 ont la même espérance, ainsi $E(M_1 - M_2) = E(M_1) - E(M_2) = \mu_1 - \mu_2 = 0$. D'où $\mu_1 = \mu_2$.*

Attention : On parle des populations, pas de l'échantillon ! Il faut considérer qu'on parle d'une vérité qu'on va ensuite démontrer (on établit la vérité dans l'hypothèse et ensuite on la démontre par le calcul à l'aide des données de l'échantillon).

I. Mise en application

La **mise en application** dépend de la **connaissance ou non de σ_1^2 et σ_2^2**

Voici un petit schéma pour vous aider :



Fiche révisions TSR – UE 4 – Le χ^2

Ces fiches ont été rédigées par des étudiants des années supérieures. Elles ne peuvent pas servir de support à une éventuelle contestation lors du concours de P.A.C.E.S.

N.B : Ce cours est un peu compliqué au premier abord. Cependant si vous le travaillez avec les exercices proposés, c'est souvent le même principe, et surtout la même démarche.

I. Principe du χ^2

Ce paragraphe peut vous aider à <u>comprendre d'où vient le Khi².</u> Ce n'est pas une application mais une explication. Il aidera notamment ceux qui ont besoin de comprendre avant tout, mais attention à ne pas vous embrouiller avec !

On définit Y, tel que $Y = Z^2$ avec $Z \sim N(0,1)$ → Cf cours sur la loi Normale.

On a donc $E(Y) = E(Z^2) = 1$. On définit aussi Zi indépendants, avec

$$Y = \sum_{i=1}^{n} Zi^2 \sim \chi^2 \text{ n ddl}$$

→ Avec n Variable aléatoire, E(Y) = n et Var(Y) = 2n

Si on additionne plusieurs lois du Khi² comme $Y_1 \sim \chi_{n1}^2$ et $Y_2 \sim \chi_{n2}^2$. On établit ainsi la loi $Y = Y_1 + Y_2 \sim \chi^2_{n1+n2}$ avec E(Y) = n1 + n2 et Var(Y) = α (n1 + n2), et ainsi le calcul de probabilités se fait par $P(Y \geq \chi^2_{\nu,\alpha}) = \alpha$

II. Comparaison d'une distribution observée à une distribution théorique

Dans ce cas on utilise le test du χ^2. Dans le formulaire vous est fournie la marche à suivre. Les différents pièges sur lesquels on peut vous attendre :

- Savoir identifier la situation (comparaison d'une distribution observée à une distribution théorique) : on vous donne un échantillon, une population, et des classes.
- Formuler les bonnes hypothèses : Cf fiche « Théorie Stat de décision – Test d'hypothèse et de signification » (pensez à parler de populations, pas d'échantillon)
- Choisir le bon test
- Faire les bons calculs
- Comparer avec le bon tableau
- Bien conclure : Cf fiche « Théorie Stat de décision – Test d'hypothèse et de signification » (Pensez que H_0 rejetée = H_1 acceptée, et pensez à ajouter le α)

En pratique ? Voici un petit schéma pour vous aider :

1. Construire un tableau avec O_i (nombre de cas observés) et P_i (proportion de cas observés). Noter l'**effectif total** et les totaux de chaque classe (π_i)

	Classe 1	Classe 2	Classe …	
O_i	O_1	O_2	O…	Effectif Total
P_i	P_1	P_2	P…	Total (= 1)
π_i	π_1	π_2	π…	

2. Formuler les hypothèses H_0 et H_1
Ex : Distributions identiques pour H_0 et Distributions différentes pour H_1

3. Ajouter les A_i (effectifs attendus) calculés par **$A_i = \pi_i$ x Effectif Total**

		Classe 1		Classe 2		Classe …		
O_i	A_i	O_1	A_1	O_2	A_2	O…	A…	Effectif Total
P_i		P_1		P_2		P…		Total (= 1)
π_i		π_1		π_2		π…		

4. Etablir la loi (k est le nombre de classe)

Ai ≥ 5

$$\rightarrow Y = \sum_{i=1}^{k} \frac{(Oi - Ai)^2}{Ai} \sim \chi^2_{k-1\ ddl}$$

↓

Faire le calcul de Y puis se référer au tableau correspondant.

Ai < 5

→ **Regrouper les classes pour arriver à des Ai ≥ 5**

N.B : si on arrive au point d'avoir 2 classes, on se retrouve avec k-1 ddl soit 1 ddl, donc une loi normale !

Puis on reprend les calculs comme à gauche

5. Conclure le test

$Y < \chi^2_{k-1,\alpha}$ → H_0 est **non rejetée**

$Y \geq \chi^2_{k-1,\alpha}$ → H_0 est **rejetée**

III. Comparaison de deux distributions observées

C'est un peu le même principe que précédemment. On utilise le même test !
Les calculs diffèrent légèrement, et surtout attention à **ne pas confondre les deux situations au départ.**

Les pièges sont les mêmes que précédemment. Voici quelques astuces :

- Dans les exercices, vous aurez deux échantillons, avec des classes similaires. Il faut que vos **classes** dans le tableau **soient les mêmes** entre les différentes lignes (quitte à faire des regroupements de classes).
- Dans le tableau ici, chaque ligne correspond à une population. Attention parfois, on inverse les colonnes et les lignes dans les exercices ! Ne vous perdez pas, et réécrivez votre tableau si c'est plus simple pour vous.

Pour la **conclusion** de cette situation :

- On conclue :
  - Rejet de H_0 (acceptation de H_1) à un risque **α inférieur à …**
  - H_0 non rejetée à un risque **α de …**
- Pour le rejet on peut ajouter :
  - Le sens du rejet (*Ex : Si on étudie le lien entre sport et obésité, et que H_0 est rejetée dans le sens qu'il y a moins d'obésité chez les gens sportifs, on peut donner ce sens dans la conclusion).*

Comment voir le sens : quand on fait la **différence entre effectif attendu et effectif observé**, regardez le signe, cela vous donnera le sens (augmentation si sens positif, et diminution si sens négatif)

De manière générale, le **α** est le **risque étudié**, souvent à 5%.

1. Construire un tableau avec O_{ij} (nombre de cas observés) et Noter l'**effectif total** et les totaux de chaque colonne (C_i) et de chaque ligne (L_j)

	Classe 1	Classe 2	Classe …	L_j
O_{i1}	O_{11}	O_{21}	O_{i1}…	L_1
O_{i2}	O_{12}	O_{22}	O_{i2}…	L_2
C_i	C_1	C_2	C…	Effectif Total

2. Formuler les hypothèses H_0 et H_1
Ex : Indépendance des distributions pour H_0 et Dépendance des distributions pour H_1
Ou encore égalité des distributions pour H_0 et inégalité (ou différence) des distributions pour H_1

3. Ajouter les A_{ij} (effectifs attendus) calculés par $A_{ij} = \frac{Ci \times Lj}{Effectif\ Total}$

	Classe 1	Classe 2	Classe …	L_j
O_{i1}	O_{11} A_{11}	O_{21} A_{21}	O_{i1}… A_{i1}	L_1
O_{i2}	O_{12} A_{12}	O_{22} A_{22}	O_{i2}… A_{i2}	L_2
C_i	C_1	C_2	C…	Effectif Total

4. Etablir la loi (k est le nombre de classe, l est le nombre de ligne)

Vérifier que $A_{ij} \geq 5$

$$\rightarrow Y = \sum_{ij} \frac{(Oij - Aij)^2}{Aij} \sim \chi^2_{(l-1)\times(k-1)}$$

Faire le calcul de Y puis se référer au **tableau correspondant**.
N.B. : (l-1)x(k-1) donne le "v" ddl (nombre de ddl)

$A_{ij} < 5$

Technique non étudiée en PACES

5. Conclure le test

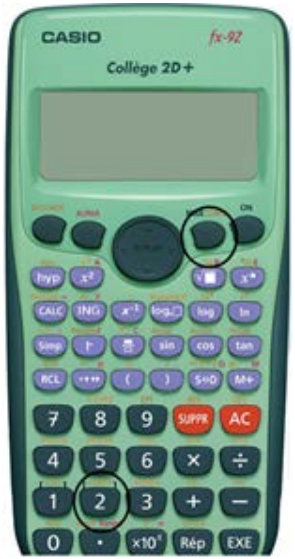
$Y < \chi^2_{v,\alpha}$ → H_0 **est non rejetée**

$Y \geq \chi^2_{v,\alpha}$ → H_0 **est rejetée**

Fiche révisions TSR – UE 4 – Utilisation de la calculatrice (Casio)

Ces fiches ont été rédigées par des étudiants des années supérieures. Elles ne peuvent pas servir de support à une éventuelle contestation lors du concours de P.A.C.E.S.

Comment utiliser la calculatrice Casio Collège 2D fx-92 pour répondre au programme de statistiques en PACES ?

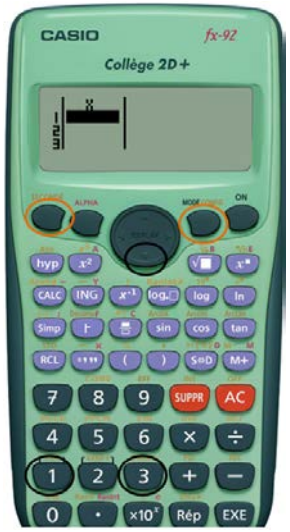


En PACES on doit utiliser à nouveau la calculatrice collège et peu de personnes ont utilisé la fonction statistique de leur calculatrice au collège, ce qui peut être utile en PACES pour éviter des calculs compliqués qui peuvent être source d'erreur.

Pour cela on va étudier le cas suivant :

Variable	25	47	78
Effectif	153	592	275

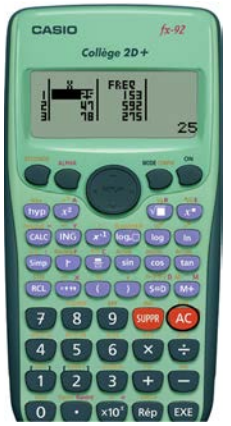
Pour commencer on va aller dans le mode statistique :
- Allez dans **"MODE"** puis allez sur **2 : "STAT"**.
- On clique ensuite sur **1 : "1-VAR"**.



A partir de là un tableau va s'afficher, sauf qu'il n'y aura qu'une colonne avec écrit X dessus, pour avoir la colonne des fréquences il faut suivre les étapes ci-dessous :
- Aller dans le mode **"SET UP"** en appuyant sur **SHIFT** puis **MODE**.
- Descendez avec la touche ↓ et allez dans le mode **3 : "STAT"**.
- La calculatrice va vous demander "Frequency?", répondre **1 : "ON"**.

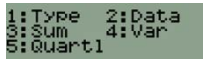
Une seconde colonne de fréquence va apparaître, à partir de là on va rentrer les valeurs dans les colonnes.

Ensuite on va retourner sur l'écran d'entrée en appuyant sur la touche **"AC"**. Si tout s'est bien passé en haut de l'écran vous verrez écrit **STAT** et non plus **Math**.



Maintenant que les valeurs sont entrées il ne reste plus qu'à rechercher les valeurs statistiques comme la moyenne ou l'écart-type.
Pour cela on va appuyer sur la touche **SHIFT** puis **1** : **"STAT"**.

A partir de là on va avoir plusieurs onglets :



- **Type** : ce n'est pas important ici.
- **Data** : ça permet de retourner au tableau de valeurs.
- **Sum** : cela permet de calculer les Σx^2 et les Σx.
- **Var** : on va y revenir plus tard.
- **MinMax** : cela sert à calculer le minimum et le maximum.

Dans l'onglet **Var** on va avoir plusieurs possibilités :

- **n** : permet de calculer la taille totale de l'échantillon.
- **$\overline{x}$** : permet de calculer la moyenne.
- **xσn** : permet de calculer l'écart-type réel (σ_n).
- **xσn-1** : permet de calculer l'écart-type estimé ($\sigma n-1$).

En appuyant sur ces symboles ils vont s'afficher sur l'écran, il suffit d'appuyer sur **"EXE"** et on obtient ces valeurs.

Au final :

Σx^2 = 3 076 453 **Σx = 53 099**	**n = 1020** **xσn = 17.4960**	**$\overline{x}$ = 52.0548** **xσn-1 = 17.5046**

Pour retourner au mode normal il suffit de réappuyer sur **MODE** puis sur **1** : **"COMP"**.
A partir de là le mot **STAT** va disparaître an haut de l'écran pour laisser place à **Math**.