Introduction aux Statistiques pour le futur Data Analyst

Alain, Stéphane et Maëlle

2019-05-27

Contents

I Introduction **3**

1 Que sont les statistiques? **3**

2 Probabilité vs Stat **4**

3 Statistiques descriptives vs statistiques inférentielles **4**

4 Statistiques vs Machine Learning **4**

II Echantillonnage **6**

1 Types d'échantillonnage **6**

2 Les types de données/variables **8**

III Estimation paramétrique **10**

1 Estimation ponctuelle **10**
1.1 Méthodes Des Moments E.M.M. - (*Method of Moments M.o.M.*) 11
1.2 La Méthode du Maximum de Vraisemblance - (*Maximum Likelihood Estimation*) 13

2 Qualité d'un estimateur **15**
2.1 Biais - *Bias* 15
2.2 Erreur quadratique (*Mean Square Error MSE*) 16

3 Méthode des Moindres Carrés **16**

4 Estimation continue: Intervalle de confiance **18**

IV Les tests paramétriques **22**

1 Les Tests d'Hypothèse **22**
1.1 Hypothèses 22
1.2 Seuil de significativité α 22
1.3 Construire l'échantillon 22
1.4 p-value & valeur-critique 23
1.5 Décision 23
1.6 Exemple: Chocolats & Stickers 23

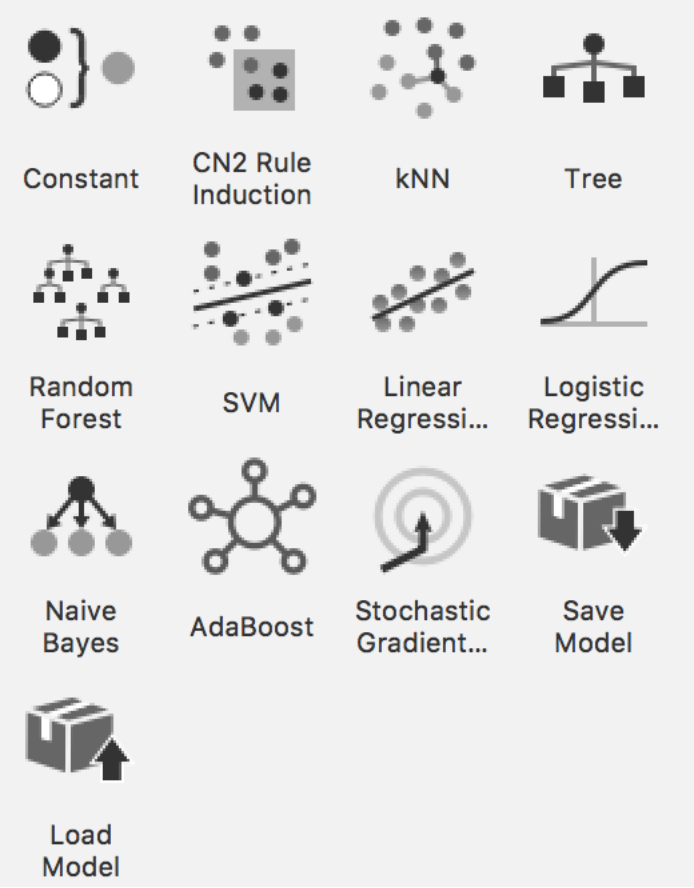
2 Test-t de Student **26**
2.1 Test de Student pour échantillon unique 26
2.2 Test-t de Student pour séries appariés 27

3 Test-f de Fisher **29**
3.1 Loi du χ^2 (Chi Deux) 29
3.2 Le test-f 29
3.3 Statistique de Test de Fisher 30
3.4 Exemple d'application 31

Part I
Introduction

1 Que sont les statistiques?

Les statistiques sont l'ensemble des techniques et méthodes permettant de collecter, organiser, analyser et interpréter des données. Population, individus, variables, données et statistiques sont les objets de "LA" statistique. Dans le domaine de l'application statistique, on parle de "modélisation statistique", c'est-à-dire que l'on appliquera des modèles statistiques à ces données afin de nous aider dans la prise de décision. On peut voir quelques exemples de modèles statistiques sur la figure suivante.



Modèles statistiques

2 Probabilité vs Stat

Probabilités et statistiques sont opposées dans leur manière d'appréhender un problème. La théorie des probabilités étudie des phénomènes caractérisés par le hasard et l'incertitude. On considère un évènement comportant un caractère aléatoire (par exemple, un lancé de dé) et on essaye ensuite de déterminer ce qui se passe.

Comme vu précédemment, les statistiques consistent à recueillir, traiter et interpréter un ensemble de données. Ainsi, on observe quelque chose qui s'est passé et essayons de comprendre quel évènement expliquerait ces observations. Il existe des interconnexions entre ces deux domaines des sciences de l'aléatoire.

Ces domaines mathématiques sont en relation avec les autres domaines mathématiques comme l'algorithmique, l'analyse, l'informatique théorique ou la logique. Les probabilités se retrouvent dans la théorie des jeux, la biologie, l'économie ou la physique, entre autres. On retrouve les statistiques dans des domaines comme l'économie, la physique, la sociologie, etc.

3 Statistiques descriptives vs statistiques inférentielles

Les statistiques descriptives fournissent un récapitulatif concis des données. On peut récapituler les données sous forme numérique ou graphique. Par exemple, le responsable d'un restaurant rapide analyse le temps d'attente des clients à l'heure du déjeuner pendant une semaine, puis il récapitule les données. Ainsi, on peut organiser, visualiser et résumer l'information recueillie.

Les statistiques inférentielles utilisent un échantillon aléatoire de données d'une population afin de décrire cette dernière et d'établir une conclusion sur les caractéristiques (ou paramètres) de cette population.

4 Statistiques vs Machine Learning

Machine Learning (ML) et statistiques sont deux domaines d'études étroitement liés. À tel point que les statisticiens parlent de machine Learning comme des statistiques appliquées ou apprentissage statistique.

Toutefois, en ML l'algorithme est capable d'apprendre à partir de données sans recourir à une programmation basée sur des règles. Il sera davan-

tage utilisé pour faire des prévisions, de l'apprentissage supervisé, ou non-supervisé.

ML et statistiques partagent donc les mêmes racines, mais diffèrent dans leurs utilisations et vocabulaire comme vu dans le tableau suivant:

Machine learning	Statistics
network, graphs	model
weights	parameters
learning	fitting
generalization	test set performance
supervised learning	regression/classification
unsupervised learning	density estimation, clustering
large grant = $1,000,000	large grant = $50,000
nice place to have a meeting: Snowbird, Utah, French Alps	nice place to have a meeting: Las Vegas in August

Différence de vocabulaire mais même signification.

Le tableau ci-dessus montre quelques différences dans le vocabulaire entre ML et statistiques. Par exemple, on parlera de modèle (ou model) en statistique alors que l'on parlera plus de réseau (ou network) en ML.

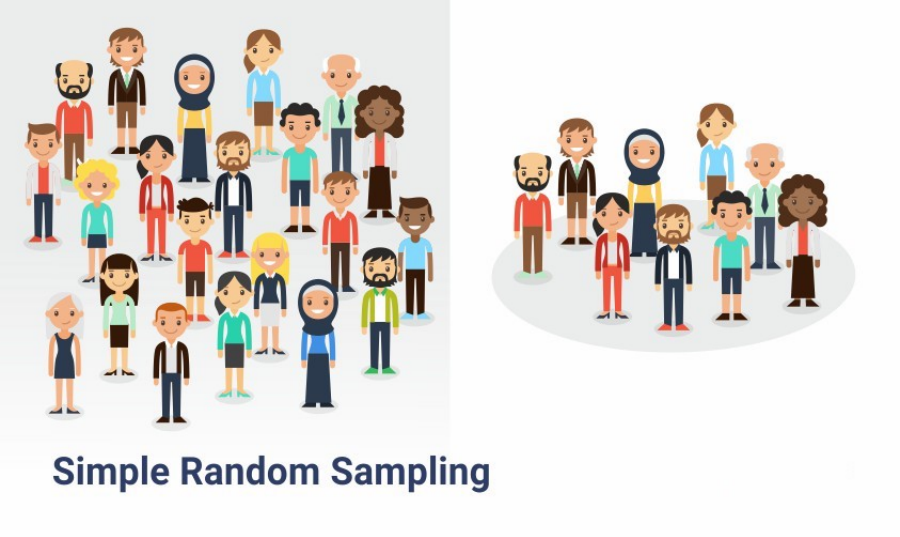
Part II
Echantillonnage

Créer un échantillon d'une population d'individus ou groupe d'éléments, c'est prélever un ou plusieurs sous-groupes qui nous permettent de représenter au mieux cette population et d'estimer certaines de ses caractéristiques.

1 Types d'échantillonnage

Il existe plusieurs type d'échantillonage, les trois plus communs sont les suivants:

Echantillonnage Aléatoire Simple (ou Simple Random Sampling (SRS)):

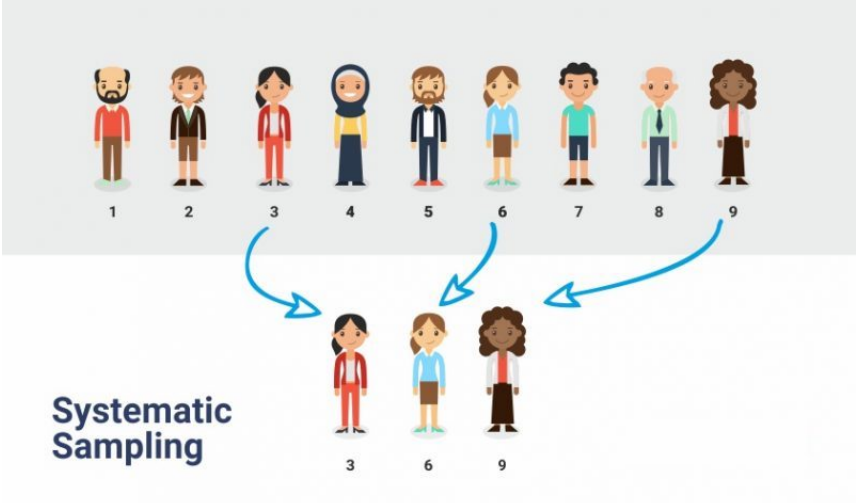


Comme le nom l'indique, les éléments ou individus d'une population sont sélectionnés simplement de manière aléatoire et chaque échantillon de taille 'n' a une chance égale d'être sélectionné.

Echantillonage stratifié (ou Stratified Random Sampling):

la population est divisée en groupe ou ”strata”. Exemple: distance maison-bureau, niveau scolaire, etc.

Echantillonnage systématiques:

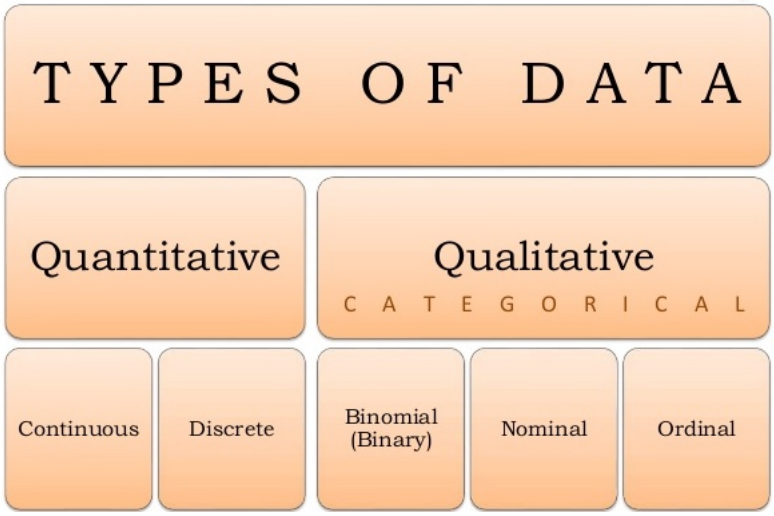


La population est ordonnée en liste, où la position du n-ième membre choisit au hasard sera répétée pour le choix du reste des membres de l'échantillon. Exemple: le 5ième membre de la liste est le premire choisit et rentre dans l'échantillon. Ainsi, les membres suivants seront sélectionnés tous les 5 pas dans la liste.

Note:
En faisant un échantillonage, il faut éviter de faire du "sous-dénombrement", où des membres de la population sont omis de l'échantillon, ce qui peut résulter en une étude biaisée ou des échantillons non-représentatifs de la population.Il est important de comprendre à quel type de données on a à faire afin d'éviter les erreurs.

2 Les types de données/variables

Il existe plusieurs types de données et le tableau ci-dessous regroupe les principaux:



Les **variables quantitatives** prennent des valeurs numériques. Toutes les opérations arithmétiques simples et complexes sont applicables aux variables quantitatives; le dénombrement (fréquences absolues), calcul de pourcentage (fréquences relatives), calcul de moyenne, médiane et écart-type jusqu'à la modélisation numérique. Ces variables se séparent en deux sous-parties: discrètes et continues.

-**Variables discrètes**: les valeurs sont distinctes et peuvent être comptées. Exemple, un dé prend 6 valeurs distinctes 1,2,3,4,5 et 6.

-**Les variables continues**: les données peuvent prendre toutes valeurs dans un interval infini ou non. Exemple: 15,0234 dans l'interval [0;80].

Les **variables qualitatives** contiennent des valeurs qui expriment une qualité, une condition, un statut unique et exclusif comme le sexe, la couleur ou bien encore la catégorie socio-professionnelle. On distinguera trois sous-groupes: les variables **binomial, nominal et ordinal**.

-**Binomial**: chaque évènement peut résulter en seulement deux valeurs possibles (ex: pile ou face, vrai ou faux),

-**Nominal**: les variables sont ”labélisées” sans valeurs quantitatives (ex: mâle femelle). Ce type de données ne peut pas être ordonné ni mesuré,

-**Ordinal**: les valeurs suivent un ordre naturel ou ”scale”.Exemple: le taux de satisfaction sur une échelle de 1 à 10. Toutefois, les différences entre les valeurs ne peuvent pas être déterminées ou est sans importance.

Part III
Estimation paramétrique

Dans ce document, nous nous préoccuperons seulement de l'estimation **paramétrique** qui est le cadre classique de la statistique. En général, grâce aux méthodes de la statistique descriptive (visualisation de l'histogramme des fréquences), il est possible de faire une hypothèse sur la nature de la loi de probabilité de notre population/échantillon. Il faut alors estimer le ou les paramètres de cette loi, puis éventuellement valider ou pas cette hypothèse.

Donc l'estimation paramétrique consiste à faire une hypothèse au préalable sur la loi de probabilité de la population.

D'un autre côté la statistique **non-paramétrique** s'applique quand on ne peut pas modéliser les observations par l'une des lois de probabilité usuelle. On ne fait donc aucune hypothèse et on essaie d'estimer directement avec précision une distribution inconnue dont les paramètres seront souvent nombreux et dont la quantité pourra varier selon le nombre d'observations (peut tendre vers l'infini). Par exemple, pour traiter des images (plus on a de pixels plus on a de paramètres à estimer: les pixels)

Dans le cadre de la statistique paramétrique on se posera deux questions:

1. Comment estime-t'on un paramètre?
2. Comment mesure-t'on la précision de ce paramètre?

Nous allons voir qu'on peut chercher à approcher directement la valeur d'un paramètre par un réel, c'est l'estimation ponctuelle, ou donner un intervalle de valeur (réelles) le plus petit possible autour du paramètre à estimer où l'estimateur en question aura de forte chance de se trouver, c'est l'estimation par intervalle de confiance.

On notera:

- θ le paramètre à estimer
- $\hat{\theta}$ ou $\hat{\theta_n}$ l'estimateur de θ (réel ou vecteur de réels si on estime plusieurs paramètres)

1 Estimation ponctuelle

Estimer un paramètre θ d'une population c'est tenté d'en approcher sa valeur en se basant sur un échantillon de cette population. Un estimateur est donc une variable aléatoire d'échantillon et l'estimation est la valeur que prend l'estimateur $\hat{\theta}$ pour cet échantillon.

Un paramètre peut avoir plusieurs estimateurs possibles. Par exemple pour estimer la moyenne d'une population on pourrait utiliser la moyenne empirique, la médiane ou encore le mode. Alors comment choisir? L'idée étant d'avoir l'estimateur qui se rapproche le plus possible du paramètre inconnu.

1.1 Méthodes Des Moments E.M.M. - (*Method of Moments M.o.M.*)

C'est la méthode la plus intuitive. Lorsqu'on estime un moyenne par une moyenne empirique ou une variance par une variance empirique on utilise la méthode des moments! Il est ainsi possible d'estimer les paramètres d'une loi à partir des moments d'un échantillon.

Un moment? Le moment d'ordre k d'une variable aléatoire X est: $E[X^k]$

- Moment d'ordre 1 = Espérance : $E[X]$
- Moment d'ordre 2 : $E[X^2]$
- Moment centré d'ordre 2 = Variance : $E[(X - E[X])^2]$

  La variance c'est aussi le moment d'ordre 2 moins le moment d'ordre 1 au carré: $Var[X] = E[X^2] - E[X]^2$

  Pour les modèles plus complexe, les moments d'ordres 3 et 4 peuvent aussi s'avérer utiles:

- Moment d'ordre centré d'ordre 3 & Skewness (cf. Figure 1)

  Skewness ou coefficient d'asymétrie: Indique à quel point la moyenne et la médiane sont décalées par rapport au mode (valeur avec la plus haute fréquence). Permet de savoir comment se différencient les valeurs extrêmes dans les queues de distribution.

  - Moment centré d'ordre 3: $\mu_3 = E[(X - E[X])^3]$
  - Skewness: $\frac{\mu_3}{\sigma^3}$

- Moment d'ordre 4 & Kurtosis (cf. Figure 2)
  Kurtosis ou coefficient d'aplatissement: il définit à quel point les valeurs extrêmes sont importantes dans la distribution. Plus les coefficients sont grands plus la courbe est aplatie par rapport à une distribution Normale

  - Moment centré d'ordre 4: $\mu_4 = E[(X - E[X])^4]$
  - Coefficient de Pearson: $\beta_2 = \frac{\mu_4}{\sigma^4}$
  - Coefficient de Fisher: $\gamma_2 = \frac{\mu_4}{\sigma^4} - 3$

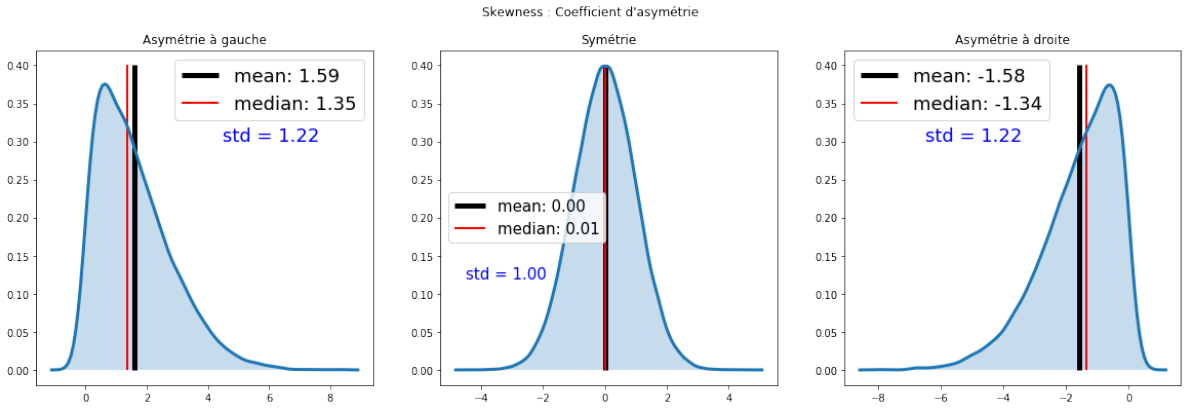


Figure 1: Skewness

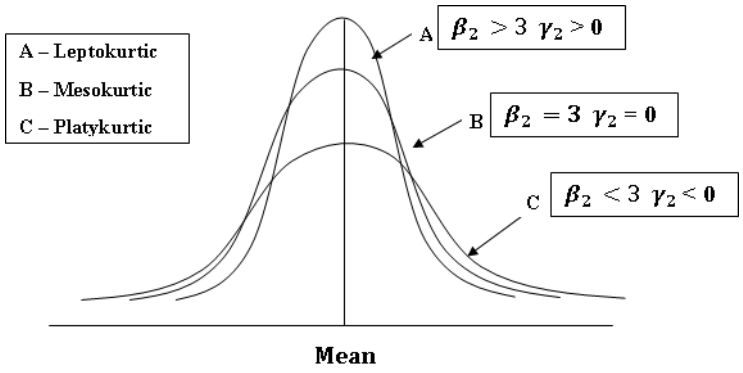


Figure 2: Kurtosis

Estimer la moyenne d'une population par l'E.M.M Si μ est la moyenne de la population qu'on cherche à estimer, on note $\hat{\mu}$ l'estimateur de μ par l'E.M.M. On a:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Estimer la variance d'une population par l'E.M.M Comme précisé précédemment, pour estimer la variance σ^2 d'une population il suffit d'estimer le moment d'ordre 2 de la population puis d'en soustraire le carré du moment d'ordre 1:

$$\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - (\frac{1}{n}\sum_{i=1}^{n} X_i)^2$$

Limites: Cette méthode peut donner une estimation qui sort de l'ensemble des valeurs possibles du paramètre. Ce problème survient en général quand la taille de l'échantillon est trop faible pour que la loi des grands nombres (cf. le magnifique article de proba) ne puisse être appliquée (les estimateurs des moments ne tendent pas vers les moments théoriques).

1.2 La Méthode du Maximum de Vraisemblance - (*Maximum Likelihood Estimation*)

Rappelons que l'on est dans le cas où on fait l'hypothèse que notre variable aléatoire suit une certaine loi de probabilité dont on cherche à en estimer le ou les paramètres inconnus θ.

Commençons par définir la vraisemblance. Il s'agit d'une fonction du paramètre θ qui mesure la probabilité que la loi (supposée) de notre population ait produit un tel échantillon.

En cherchant à la maximiser on obtient alors l'estimation la plus vraisemblable pour θ, c'est-à-dire la plus plausible ou la plus probable pour qu'on ait eu précisément cette échantillon depuis la loi.

Autrement dit, il faut trouver le(s) paramètre(s) qui maximise(nt) la probabilité d'avoir obtenu les observées (celle de l'échantillon) à partir de ce modèle en particulier.

C'est la probabilité d'avoir θ sachant $X_1 = x_1, X_2 = x_2, .., X_n = x_n$

$$\begin{aligned} L(\theta|X_1, .., X_n) &= \text{probabilité jointe de } X_1, .., X_n\theta \\ &= P(X_1 = x_1, .., X_n = x_n|\theta) \\ &= \prod_{i=1}^{n} P(X_i = x_i|\theta) \end{aligned} \tag{1}$$

Exemple Prenons une population de taille N avec une proportion p de femme et $1-p$ d'homme. On cherche à estimer le paramètre p, proportion de femme dans toute la population. Pour cela on tire un n-échantillon aléatoire i.i.d. et on définit X_i la variable aléatoire telle que:

$$X_i = \begin{cases} 1 & \text{si femme} \\ 0 & \text{si homme} \end{cases}$$

La fonction de répartition des X_i est la suivante (Loi de Bernouilli):

$$F(X_i = x_i|p) = p^{x_i}(1-p)^{1-x_i}$$

Ainsi, pour une observation on a:

- si $X_i = 1 => F(X_i|\theta) = p$

- si $X_i = 0 => F(X_i|\theta) = 1 - p$

Et pour tout l'échantillon:

$$L(p|X_1 = x_1, X_2 = x_2, .., X_n = x_n) = p^{x_1}(1-p)^{1-x_1} * p^{x_2}(1-p)^{1-x_2} * .. * \hat{p}^{x_n}(1-p)^{1-x_n}$$
$$= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} \tag{2}$$

Pour trouver le maximum de cette fonction il faut:

1. La dériver par rapport à p
2. Trouver la valeur de p pour laquelle la dérivée s'annule

Notons que dériver un produit de fonctions est laborieux, c'est pour cela qu'on préfère maximiser le logarithme de la fonction à maximiser. En effet, passer par la fonction logarithme à l'avantage de nous faire passer d'un produit de fonctions à une somme de fonctions, beaucoup plus simple à dériver, tout en nous garantissant que le max sera atteint au même point (monotonie et croissance du logarithme népérien).

Rappel:

$$\ln ab = \ln a + \ln b$$
$$\ln a^b = b \ln a$$

Ainsi

$$\ln[\prod f(x_i)] = \sum \ln f(x_i)$$

notons ℓ, la fonction telle que $\ell = \ln L$

$$\ell = \ln \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$$
$$= \sum_{i=1}^{n} \ln(p^{x_i}(1-p)^{1-x_i})$$
$$= \sum_{i=1}^{n} X_i \ln(p) + X_i \ln(1-p)$$
$$= \ln(\hat{p}) \sum_{i=1}^{n} x_i + \ln(1-p) \sum_{i=1}^{n} 1 - x_i$$

On identifie:

$$\sum_{i=1}^{n} X_i = n\bar{X}$$
$$\sum_{i=1}^{n} 1 - X_i = n(1-\bar{X})$$

Il en résulte:

$$\ell = n\bar{X} \ln p + n(1 - \bar{X}) \ln(1 - p)$$

On cherche $\hat{p}$ qui annule la dérivé partielle de ℓ en $\hat{p}$:

$$\frac{\partial \ell}{\partial \hat{p}} = \frac{n\bar{X}}{\hat{p}} - \frac{n(1 - \bar{X})}{1 - \hat{p}} = 0$$
$$<=> \hat{p} = \bar{X}$$

Comme on pouvait s'y attendre l'estimateur du maximum de vraisemblance est tout simplement la proportion de femme dans l'echantillon.

Notons que l'EMV est le meilleur estimateur non biaisé (*Best Unbiased Estimator*)

2 Qualité d'un estimateur

2.1 Biais - *Bias*

Le biais c'est tout simplement la différence entre l'espérance de notre estimateur $\hat{\theta_n}$ et la vraie valeur du paramètre estimé θ:

$$Biais = E[\hat{\theta_n}] - \theta$$

Estimateur non-biaisé: Si le biais de l'estimateur est nul cela signifie qu'en moyenne notre estimateur est égal au paramètre θ. Et donc en moyenne on ne se trompe pas.

On parle d'estimateur **asymptotiquement non-biaisé** quand un estimateur est a priori biaisé mais plus la taille de l'échantillon est grande plus ce biais est négligeable.

$$\lim_{n \to \infty} E[\hat{\theta_n}] - \theta = 0$$

Biais pour la moyenne empirique

$$E[\bar{X}] = E[\frac{1}{n} \sum_{i=1}^{n} X_i]$$

$$E[\bar{X}] = E[X] = \mu$$

Le moyenne empirique est un estimateur non biaisé de la moyenne théorique

Biais de la variance empirique

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i - \bar{X_n}^2$$
$$E[\hat{\sigma}^2] = \frac{n-1}{n}\sigma^2$$

La variance empirique est biaisée mais asymptotiquement sans biais:

Biais de la variance empirique corrigée $\hat{\sigma*}2$

$$\hat{\sigma}^{*2} = \frac{n}{n-1}\hat{\sigma}^2$$
$$E[\hat{\sigma}^{*2}] = \sigma^2$$

La variance empirique corrigé est non-biaisée:

$$\lim_{n\to\infty} E[\hat{\sigma}^{*2}] - \sigma = 0$$

2.2 Erreur quadratique (*Mean Square Error MSE*)

L'erreur quadratique est une mesure de dispersion tout comme la variance, à la différence près qu'elle nous donne l'écart moyen entre l'estimateur et le paramètre à estimer.

Autrement dit, remis dans le contexte du Machine Learning, il représenterais la moyenne des écarts entre la prévision du modèle et nos observations.

Et c'est cette valeur qu'on cherche à minimiser lors d'une régression linéaire (simple ou multiple) (une semaine y sera consacrée très prochainement).

$$Erreur\ Quadratique = E[(\hat{\theta_n} - \theta)^2]$$

On peut aussi l'écrire de cette façon:

$$Erreur\ Quadratique = Biais(\hat{\theta})^2 + Var[\hat{\theta}]$$

On cherche en général à trouver l'estimateur qui minimise cette erreur. Il s'agit souvent de trouver un compromis entre un biais petit ou nul et une petite variance. C'est ce qui est fait lors d'une régression linéaire. On cherche la droite qui minimise les erreurs quadratiques.

3 Méthode des Moindres Carrés

La méthode des moindres carrés, permet de comparer des données expérimentales, généralement entachées d'erreurs de mesure, à un modèle mathématique censé décrire ces données. Pour trouver l'équation de la droite on devra chercher la somme minimale des distances des points par rapport à la

droite.

Par exemple :

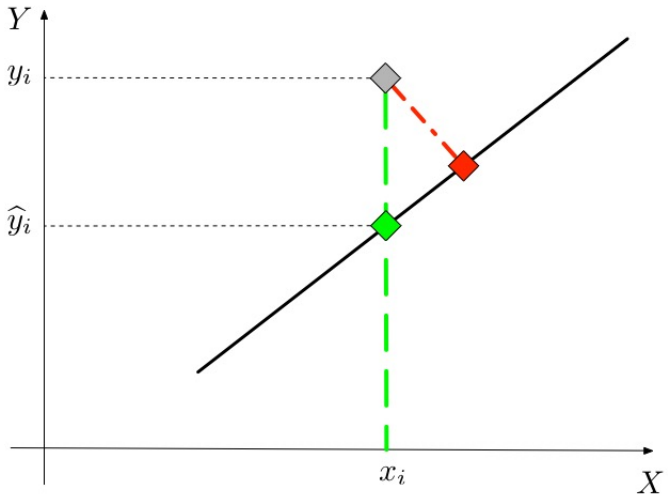


Figure 3: distance des erreurs

Comment mesurer cette distance ?

Une situation courante en sciences est d'avoir à sa disposition deux ensembles de données de taille n, y1, y2,...,yn et x1, x2,...,xn, obtenus expérimenalement ou mesurés sur une population. Le problème de la régression consiste à rechercher une relation pouvant éventuellement exister entre les x et les y, par exemple de la forme y = f(x). Lorsque la relation recherchée est affine, c'est-'a-dire de la forme y = ax + b, on parle de régression linéaire. Mais même si une telle relation est effectivement présente, les données mesurées ne vérifient pas en général cette relation exactement. Pour tenir compte dans le modèle mathématique des erreurs observées, on consid'ere les données y1, y2,...,yn comme autant de réalisations d'une variable aléatoire Y et parfois aussi les données x1, x2,...,xn comme autant de réalisations d'une variable aléatoire X. On dit que la variable Y est la variable dépendante ou variable expliquée et que la variable X est la variable explicative .

Finalement on est ramené a calculer les distances orientées, on compare la valeur calculée à la valeur observée en calculant l'écart entre ces deux valeurs. Si le point est au dessus de la droite l'écart aura une valeur négative, s'il est en dessous, l'écart aura une valeur positive.

Donc nous devons calculer la somme au carrés des erreurs.

$$SCE = \sum_{i=1}^{n} (y_i - (ax_i + b)^2$$

Ce calcul nous rammène a deux équation qui vont nous permmetre de résoudre l'équation de la droite passant par nos données.

$$y = ax + b$$

$$a = \frac{\bar{x}\bar{y} - \bar{x}.\bar{y}}{\bar{x^2} - \bar{x}^2}$$

$$b = \bar{y} - a.\bar{x}$$

Cela nous permet d'obetenir cette droite pour l'ensemble de nos donées

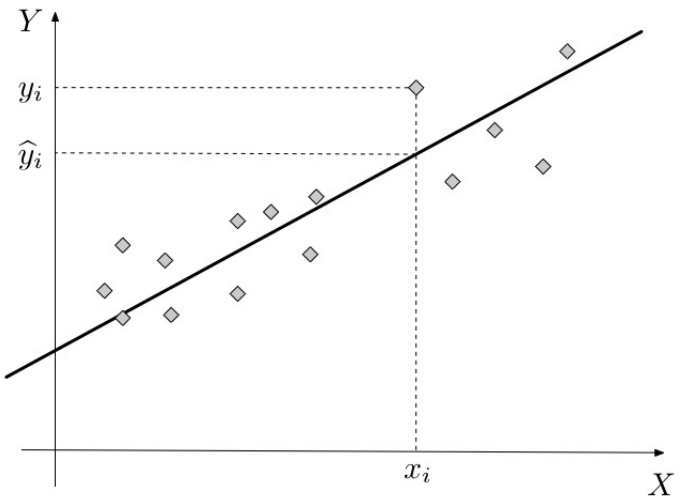


Figure 4: Droite des moindres carrés

4 Estimation continue: Intervalle de confiance

Les intervalles de confiance de certaines statistiques spécifiques (par exemple, les moyennes, ou droites de régression) donnent un intervalle de valeurs autour du paramètre où l'on peut s'attendre à ce que la "véritable" statistique (de la population) se situe (avec un certain degré de certitude) Vous êtes un agriculteur et cette année votre récolte est de plus de 200 000

pommes .Vous prélevez un échantillons de 36 pommes . Dans cette récolte le poids moyen d'une pomme est de 112 grammes avec un écart type de 40 g . Quelle est la probabilité que le poids moyen de nos 200 000 pommes soit entre 100 et 124g

- $\bar{X}$ = moyenne d'un échantillon
- mu= espérance
- σ = écart type

Nous savons que Le théorème centrale limite (aussi improprement appelé théorème de la limite centrale ou centrée) établit la convergence en loi de la somme d'une suite de variables aléatoires vers la loi normale. Intuitivement, ce résultat affirme que toute somme de variables aléatoires indépendantes tend dans certains cas vers une variable aléatoire gaussienne.
Donc comme nous allons tirez successivement plusieur échantillons de moyenne μ_x et d'écart type σ_x
nous allons congerger ver une loi normale :

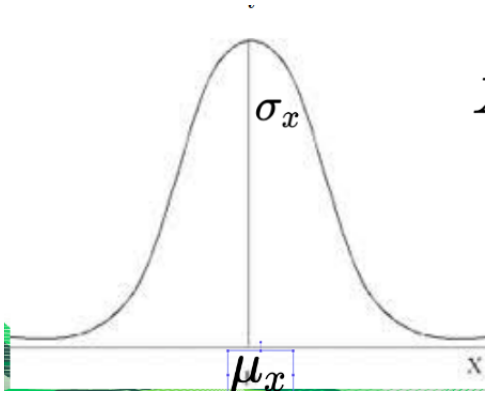


Figure 5: Distribution des echantillons

Nous voyons bien que nos échantillons tendent ver une loi normale.
Ce que nous voulons savoir c'est quelle proportion de cette distribution tombe entre 100 et 124 g
Pour cela analysons notre distribution.
Nous savons que:

$$\sigma_x = \frac{\sigma}{\sqrt{n}}$$
$$\frac{\sigma}{6} = \frac{40}{6}$$
$$\mu_x = 112$$

Donc ce que nous voulons c'est la probabilité que (mu) la myenne de nos échantillons soit entre 100 et 124

calculons donc la probabilité :

$$P(\mu - 12 < \bar{X} < \mu + 12)$$

$$P(-12 < \bar{X} - \mu_x < \mu + 12)$$

Pour avoir ma borne(de mon intervalle) je vais exprimer ce résultat en fonction de sigma

$$"Borne" = \frac{12}{40/6} = 1.8$$

$$P(-1.8 < \frac{Mx - \mu_x}{\sigma_x} < 1.8)$$

Nous voyons que la probabilité que la moyenne des échantillons soit entre 100 et 124g et de 1.8g . comme nous avons exprimer ce résultat en fonction de sigma nous nous reportons à la table de la loi normale centré réduite .. Que vous montre ci dessous
Nous y trouveront 1.8 comme prévue .

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	[illegible]	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	[illegible]	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Figure 6: Table loi normale

Nous voyons que 1.8 dans la table de la lois normale = 0.9641
Cela veut dire que tout ce qui est derriere 1.8 = 96,41%
Étant donné de que veut tout les élément à 1.8 écart type de ma moyenne(tout ce qui est hachuré) .
Fonction de densité de la loi normale centrée réduite dis que l'aire en dessous de la courbe = 1 , qu'elle est centré en 0 et avec un écart type de 1
Ma figure est centré en 0 donc l'aire avant == 0.5 et l'aire après == 0.5.
Donc 0.9641 - 0.5 et je trouve la partie violette qui est à 1.8 écart type de

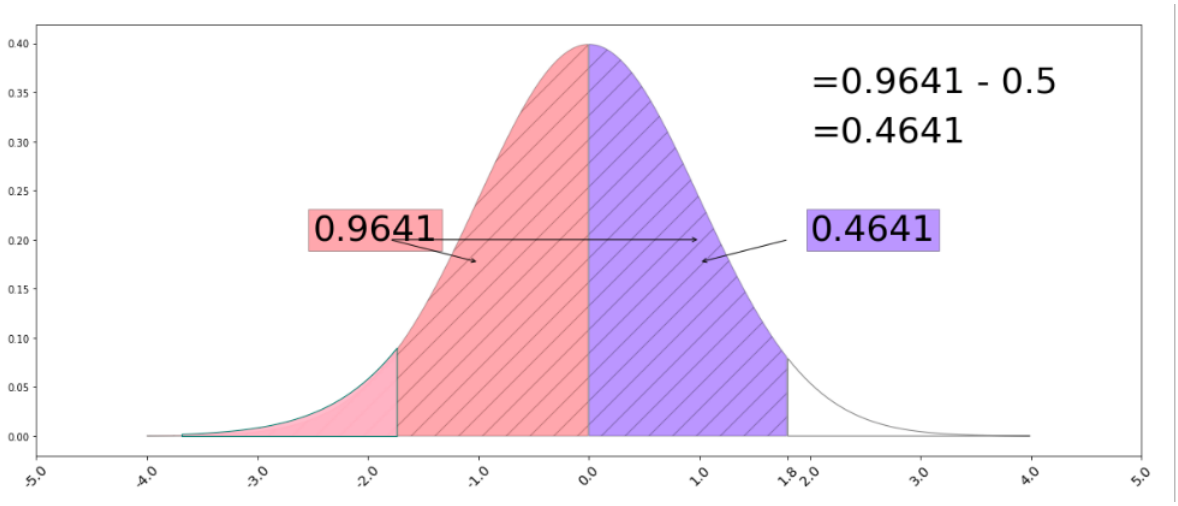


Figure 7: loi normale centré reduite

ma moyenne .
Donc 0.9641-0.5 = 0.4641
P = 2 * 0.4641 = 0.9282 = 92,82
J'ai donc une probabilité de 92.8 % que la moyenne de notre récolte soit entre 100 et 124 G.

$$P(-1.8 < \frac{\bar{X} - \mu}{\sigma} < 1.8) = 92.82\%$$

Part IV
Les tests paramétriques

1 Les Tests d'Hypothèse

Les tests d'hypothèse sont un autre aspect important de l'inférence statistique. La démarche est la suivante:

1. On étudie une population dont les éléments suivent une loi usuelle dont le(s) paramètre(s) sont inconnu(s)
2. On émet une hypothèse sur ce(s) paramètre(s)
3. On teste cette hypothèse sur un échantillon prélevé de cette population
4. On décide de rejeter ou non cette théorie, avec un faible risque de se tromper.

1.1 Hypothèses

- **Hypothèse nulle** H_0: c'est l'hypothèse de départ, celle qui sera supposé vraie durant les tests et qu'on voudra contredire.
- **Hypothèse alternative** H_1: c'est ce qu'on essaie réellement de prouver. Elle doit contredire H_0

Pourquoi préférer définir ces hypothèse de sorte à ce qu'on essaie de prouver que H_0 est fausse? Simplement parce qu'il est souvent plus aisé de prouver qu'une théorie est fausse (Tous les signes sont-ils blancs? En trouvant un seul signe noir on démontre que c'est faux, inutile de chercher à vérifier la couleur de chaque signe!)

1.2 Seuil de significativité α

Il définie une zone de rejet de l'hypothèse nulle H_0. C'est la probabilité de rejeter H_0 alors qu'elle est vraie (erreur de type 1). On le fixe en général à 5% ou 1%.

Hypothèses, Décision et Erreurs		
	H_0 vraie	H_1 vraie
Accepter H_0	OK	Erreur de type 2 β
Rejeter H_0	Erreur de type 1 α	OK

1.3 Construire l'échantillon

Le test nous dira si l'hypothèse formulée est supportée ou non par les observations.

1.4 p-value & valeur-critique

La p-value et la valeur critique permettent de conclure si les différences observées entre la valeur de la statistique d'échantillon et celle du paramètre sur lequel l'hypothèse est faite est trop important pour être uniquement imputable au hasard de l'échantillonnage.

- **p-value**: probabilité que la loi de probabilité ait produit notre échantillon sachant que H_0 est vraie. Toute observation contredisant l'hypothèses nulle est due aux hasard.

$$p\ value = P[X_1 = x_1, .., X_n = x_n | H_0\ vraie]$$

- **valeur critique**: valeur de la statistique qui correspond à la probabilité α

$$\alpha = P[T^* < Valeur\ Critique | H_0\ vraie]$$

1.5 Décision

- si $p_{value} < \alpha$ ou $|T^*| < Valeur\ Critique$: On rejette H_0
- si $p_{value} > \alpha$ ou $|T^*| > Valeur\ Critique$: On ne rejette pas H_0
  Cela signifie que l'échantillon ne donne pas de preuves suffisantes pour rejeter H_0

Quelque soit la décision qui est prise, il faut se souvenir qu'il y a toujours un risque (α ou β) de s'être tromper, le hasard nous aura induit en erreur.

1.6 Exemple: Chocolats & Stickers

Stef vend des chocolats. Alain lui dit qu'il peut augmenter ses ventes en offrant un petit cadeaux avec chaque boîte de chocolats. Stef est sceptique, il n'a pas l'extraordinaire sens du marketing d'Alain. Alain lui propose de faire une expérience: pendant un mois, Stef offrira un cadeau avec ses chocolats certains jours et d'autres non. Il décide d'offrir des stickers. Chaque jour il tire à pile ou face pour savoir s'il offrira ou non des stickers ce jour-là.

Hypothèses

- H_0: aucune différence entre les ventes journalières moyennes avec et sans stickers offerts

$$H_0 : \mu_{sticker} = \mu_{no-sticker}$$
$$H_0 : \mu_{sticker} - \mu_{no-sticker} = 0$$

- H_1: différence significative entre les ventes journalières moyennes avec et sans stickers offerts

$$H_1 : \mu_{sticker} \neq \mu_{no-sticker}$$
$$H_1 : \mu_{sticker} - \mu_{no-sticker} \neq 0$$

N.B.: on décrit ici un test de Student (voir la section sur les tests de Student) sur deux échantillons indépendants.

Seul de significativité $\alpha = 5\%$

Echantillon 23 jours de vente dont 10 jours sans stickers offert.
On obtient les statistiques suivantes pour notre échantillon:

	Sticker	No_Sticker
count	13.00	10.00
mean	301.94	265.83
std	29.72	39.15
min	248.30	204.00
50%	295.00	279.75
max	354.80	320.60

On pourrait déjà noter que lorsque des stickers sont offerts avec les boites de chocolat, les ventes sont en moyennes plus élevées mais aussi plus stables.

```
1  import pandas as pd
2  import numpy as np
3
4  choco = pd.DataFrame(data={
5      'Sticker':[248.3, 295,338, 285.3,287.1,283.6,354.8,291.9,
6          309.8,304.6,322.2,269.4,335.2],
7      'No_Sticker':[215.1,300,320.6,276.6,282.9,288.1,220.8,283,
8          204,267.2, None, None,None]})
9  choco.describe(percentiles=[]).round(2)
```

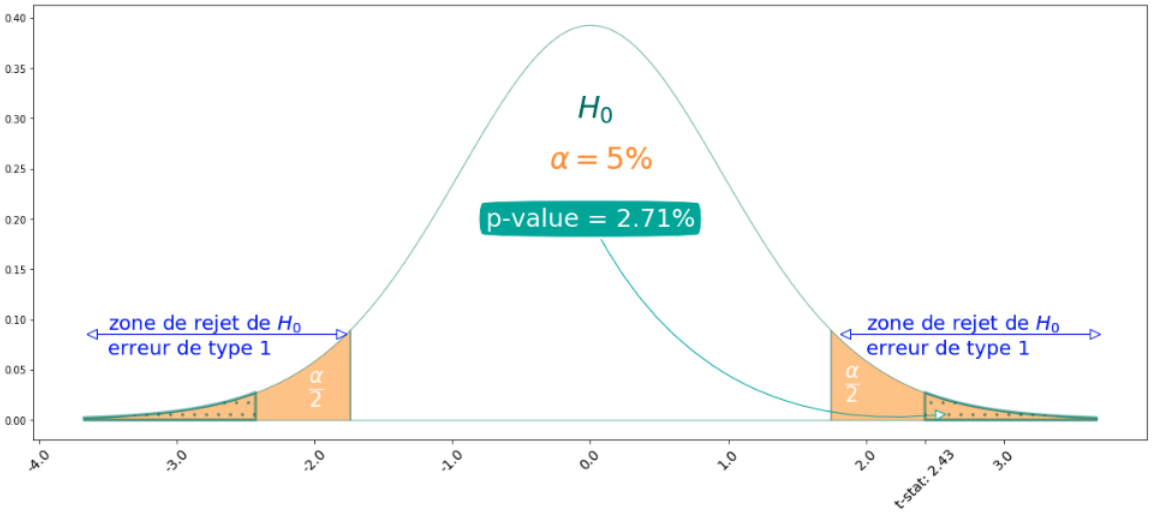
p-value = 2.7%

```
import numpy as np
from scipy import stats

r = stats.ttest_ind(choco.Sticker, choco.No_Sticker,
    equal_var=False, nan_policy='omit')
print(f'statistic {r[0].round(2)} / p-value  {r[1]:.2%}')
```

Décision $p\ value(2.7\%) < \alpha(5\%)$

On rejette H_0 car l'échantillon nous montre qu'il est peu probable d'avoir des ventes en moyennes égales qu'on offre ou non des stickers. Ce qui confirme bien l'intuition d'Alain!



2 Test-t de Student

Le test-t de Student est un test statistique permettant de comparer les moyennes de deux groupes d'échantillons. Il s'agit donc de savoir si les moyennes des deux groupes sont significativement différentes au point de vue statistique. Il existe plusieurs variants du test-t de Student:

- Le test-t de Student pour échantillon unique
- Le test-t de Student comparant deux groupes d'échantillons indépendants (on parle de test de Student non apparié)
- Le test-t de Student comparant deux groupes d'échantillons dépendants (on parle de test de Student apparié).

2.1 Test de Student pour échantillon unique

Si notre échantillons est assez grand = loi normale centrée réduite
Si notre échantillons est faible = table t de Student
Imagions que nous avons administré un médicament à 7 patients pendants 3 mois on mesure leur pression sanguine :
Distribution = [1.5 / 2.9 / 0.9 / 3.9 / 3.2 /2.1 / 1.9]
Construire un intervalle de confiance a 95 pourcent de la variation réel de la population
Comme nous sommes dans un cas ou l'échantillon es petit nous nous dirigeons ver une loi de Student
On cherche dans la table de student la valeur qui prend 95% des résultat
n = échantillons
Puis nous avons n- 1 de Degré De Lberté* ¿ n-1 = 7-1 donc 6
On obtient 2.447 Donc si on cherche un intervalles autour de la moyenne délimité à 95 il sera égale à 2,447 écart type :

$$IC = [\mu_x - 2.447\sigma_x; \mu_x + 2.447\sigma_x]$$

Maintenant que nous avons les informations pour notre intervalle . nous allons calculer σ

$$\sigma_x = \frac{\sigma}{\sqrt{n}} = \frac{1.04}{\sqrt{7}} = 0.3930830519$$

Donc :

$$2.447\sigma_x = 0.96$$

$$IC = [\mu_x - 0.96; \mu_x + 0.96]$$

$$IC = [1.38; 3,33]$$

$$IC = [1.38; 3.33]$$

je viens donc de démontrer grace à la table de Student que mon intervalle est entre 1.38 et 3.33 . Pour être plus précis je vous affiche un graphique pour que ce sois plus explicite

$$IC = [\mu_x - 2.447\sigma_x; \mu_x + 2.447\sigma_x]$$

$$IC = [2.34 - 0.96; 2.34 + 0.96]$$

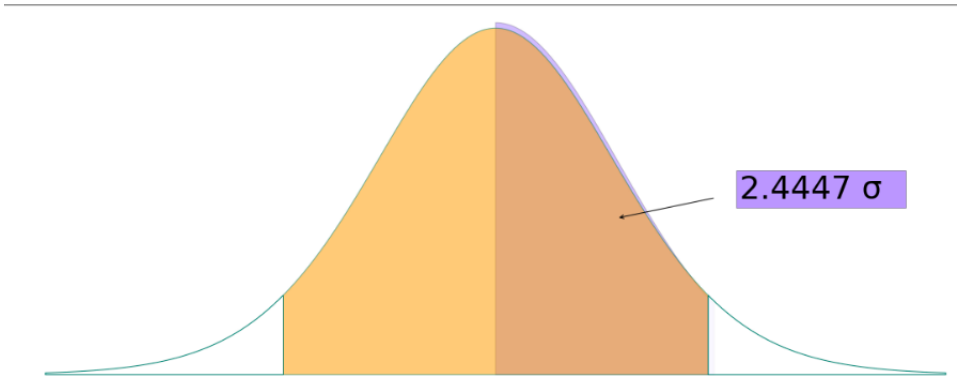


Figure 8: représentation Student

$$IC = [1.38; 3.33]$$

Nous pouvons implémenter le calcul du test t de Student pour un échantillons unique directement en Python.

```
from scipy import stats
np.random.seed(7654567)  # fix seed to get the same result
rvs = stats.norm.rvs(loc=5, scale=10, size=(50,2))
stats.ttest_1samp(rvs,5.0)
(array([-0.68014479, -0.04323899]), array([ 0.49961383,  0.96568674]))
stats.ttest_1samp(rvs,0.0)
(array([ 2.77025808,  4.11038784]), array([ 0.00789095,  0.00014999]))
```

Nous avons en résultat t-statistic et la p-value

2.2 Test-t de Student pour séries appariés

Le test de Student apparié permet de comparer la moyenne de deux séries de valeurs ayant un lien.

Par exemple, 20 souris ont reçu un traitement X pendant 3 mois. On se pose la question à savoir si le traitement X a un impact sur le poids des souris au bout des 3 mois. Le poids des 20 souris a donc été mesuré avant et après traitement. Ce qui nous donne 20 séries de valeurs avant traitement et 20 autres séries de valeurs après traitement provenant de la mesure du poids des mêmes souris.
Il s'agit bien dans cet exemple, d'un test de Student apparié car les deux séries de valeurs ont un lien (les souris). Pour chaque souris, on a deux mesures (l'une avant et l'autre après traitement).

Formule :
Pour comparer les moyennes de deux séries appariées, on calcule tout d'abord la différence des deux mesures pour chaque paire.
Soit d la série des valeurs correspondant aux différences des mesures entre les paires de valeurs. La moyenne de la différence d est comparée à la valeur 0.

S'il y a une différence significative entre les deux séries appariées, la moyenne de d devrait être très éloignée de la valeur 0
La valeur t de Student est donnée par la formule :

$$t = \frac{m}{s/\sqrt{n}}$$

m et s représentent la moyenne et l'écart-type de la différence d. n est la taille de la série d.
Pour savoir si la différence est significative, il faut tout d'abord lire dans la table t, la valeur critique correspondant au risque alpha = 5% pour un degré de liberté :
- d.d.l=n1

Nous pouvons implémenter le calcul du test t de Student couplé directement en Python.

```
1 from scipy import stats
2 np.random.seed(12345678)
3 #Test avec  chantillon  avec des moyennes identiques:
4 rvs1 = stats.norm.rvs(loc=5,scale=10,size=500)
5 rvs2 = stats.norm.rvs(loc=5,scale=10,size=500)
6 stats.ttest_ind(rvs1,rvs2)
7 (0.26833823296239279, 0.78849443369564776)
8 stats.ttest_ind(rvs1,rvs2, equal_var = False)
9 (0.26833823296239279, 0.78849452749500748)
```

Nous avons en résultat la t-statistic et la p-value

3 Test-f de Fisher

Le test de Fisher, aussi appelé **"test F d'égalité de deux variances"**, est un test d'hypothèse qui permet de tester l'hypothèse que deux variables aléatoires suivant des lois normales ont la même variance.

Le test compare une valeur dite "observée", qui est simplement le quotient des variances de deux échantillons, à une valeur dite "théorique" (ou "critique"), que l'on aura obtenue à l'aide de la table de la loi de Fisher.

3.1 Loi du χ^2 (Chi Deux)

Une autre loi, celle du **chi carré** χ^2, est également importante à savoir.
En 1900, le statisticien britannique Karl Pearson introduit une loi statistique résultant d'une interrogation:
"le hasard seul est-il responsable des écarts observés par rapport à la moyenne attendue?"

La loi du χ^2 stipule que si la distribution d'une variable aléatoire suit une loi normale de moyenne mu μ et d'écart-type sigma σ, on considérera alors que la somme des distances X_j à la moyenne $\bar{X}$ au carré par rapport à la moyenne suit une loi dite du chi carré χ^2 à théta θ degré de libertés.

$$Y = \sum_{j=1}^{J} \frac{(X_j - \bar{X})^2}{\bar{X}}] \sim \chi^2(\theta)$$

3.2 Le test-f

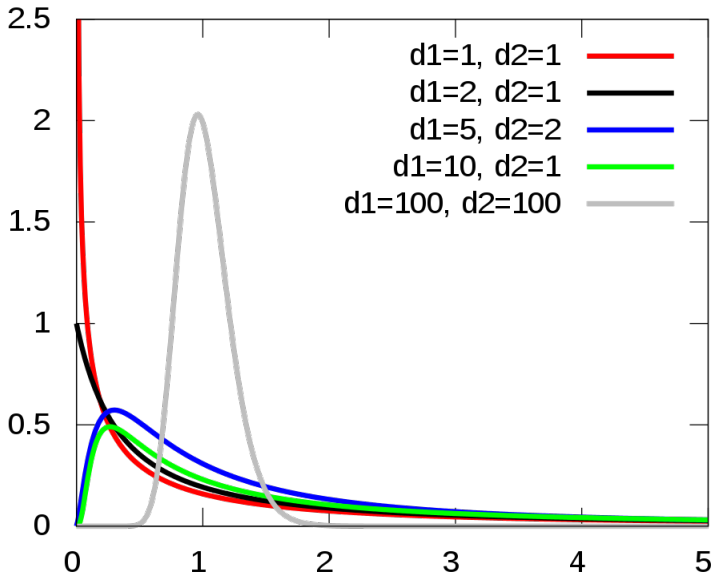
Revenons-en maintenant à ce fameux test F de Fisher.
Pour deux échantillons A et B de taille n_1 et n_2:

A = $\{X_{1,1}, ..., X_{n,1}\}$, suit une loi du Chi 2 à v_1 degré de liberté (ou ddl);
A $\sim\chi^2(v_1)$,
et B = $\{X_{1,2},...,X_{n,2}\}$ suit une loi du Chi 2 à v_2 degré de liberté:
B $\sim\chi^2(v_2)$, on aura:

$$F_{observé} = \frac{S_{n1}^2}{S_{n2}^2} \sim \ F(v_1 = n_1 - 1, v_2 = n_2 - 1)$$

Le rapport des variances de mes deux échantillons est distribué selon une loi de Fisher dont les deux nombres de degrés de liberté v1 et v2 sont la taille des échantillons n1-1 et n2-1.

Il est important de noter que plus la valeur des degrés de libertés augmentent, plus la courbe tendra vers la forme en cloche de la loi normale (ou "bell curve") pour une moyenne de "1", comme vu sur le graphique ci-dessous:



3.3 Statistique de Test de Fisher

Il nous faut maintenant la valeur $F_{théorique}$ pour comparer notre valeur observée. Dans la table de la loi de Fisher, cette valeur critique s'obtient à l'intersection des deux degrés de liberté **v1** et **v2**.

ν_2 \ ν_1	1	2	3	4	5	6	8	10	15	20	30	∞
1	161	200	216	225	230	234	239	242	246	248	250	254
2	18,5	19,0	19,2	19,2	19,3	19,3	19,4	19,4	19,4	19,4	19,5	19,5
3	10,1	9,55	9,28	9,12	9,01	8,94	8,85	8,79	8,70	8,66	8,62	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,04	5,96	5,86	5,80	5,75	5,63

Cet extrait de tableau montre les colonnes v1 et les lignes v2 d'une table de la loi de Fisher pour un seuil de signification de 5%.

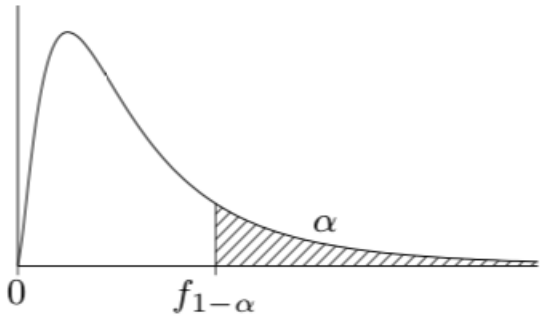
À l'intersection de v_1 et v_2, se trouve notre $F_{théorique}$ $(f_{1-\alpha})$ telle que:

$$P(F > f_{1-\alpha}) = \alpha = 0.05$$

Autrement dit, on trouvera $f_{1-\alpha}$ telle que la probabilité que $F_{observé}$ soit supérieur à $F_{théorique}$ est de 5%.

A quoi va nous servir cette statistique?

Elle nous sert de point de repère pour séparer deux zones, une zone de rejet de l'hypothèse nulle ($F_{obs} > F_{th}$) et une autre de non-rejet ($F_{obs} < F_{th}$).



Le graphique ci-dessus montre la zone de rejet de l'hypothèse nulle (pour $F > f_{1-\alpha}$*) et la zone grisée de non-rejet de l'hypothèse nulle (pour* $F < f_{1-\alpha}$*).*

Une fois que l'on a la valeur observée et la valeur critique, l'appréciation des résultats se fait de la manière suivante pour:

$$Hypothèse\ nulle\ H0 : A\ et\ B\ ont\ des\ variances\ similaires, VA = VB$$

$$Hypothèse\ alternative\ H1 : Les\ variances\ de\ A\ et\ B\ sont\ significativement\ différentes, VA \neq VB$$

.
Ainsi:
Si $F_{observée}$ est plus grande que $F_{théorique}$ $f_{1-\alpha}$: on rejettera l'hypothèse nulle H_0 que VA = VB et on se tournera alors vers l'hypothèse alternative H_1.
Si $F_{observé}$ est plus petite que le $F_{théorique}$: on ne rejettera pas l'hypothèse nulle H_0.

3.4 Exemple d'application

Un laboratoire veut tester l'efficacité d'un médicament sur la réduction du taux de cholestérol. Le laboratoire mesure le taux de cholestérol d'un groupe témoin A qui ne reçoit pas le médicament et d'autre groupe B après traitement. Le test de Fisher peut ainsi être utilisé pour comparer la variance du taux de cholestérol de chaque groupe. Si les variances des groupes

sont significativement différentes, on pourra conclure que le médicament a un effet sur les patients du groupe B.

Représentons ça en **Python**:

```
import random; import scipy; from scipy import stats; import numpy as np
#200 milligrammes par decilitre (mg/dL)  est la norme du taux de cholesterol pour un adulte
#au-dessus de 240 mg/dL le taux est trop eleve
a = [] #groupe test
b = [] #groupe temoin
for i in range(0,10): x = random.randint(200,239); a.append(x)
for j in range(0,10): y = random.randint(240, 300); b.append(y)

if np.var(a) > np.var(b): F = np.var(a) / np.var(b) #variance max / variance min;
else: F = np.var(b) / np.var(a)
F = round(F,2) #arrondi à deux chiffres apres la virgule

dfn = len(a) - 1; dfd = len(b) - 1 #les degrés de liberté: dfn = 9 for a; et dfd = 9 for b
alpha = 0.05 #le risque de faussement rejeter H0
F_theorique = scipy.stats.f.ppf(q=1-alpha, dfn=dfn, dfd=dfd); F_theorique = round(F_theorique,

print("Taux de cholestérol dans a =", a); print("Taux de cholestérol dans b =", b)
print("F_observé =" , F); print("F_theorique =", F_theorique)

if F > F_theorique: print("reject H0"),
else: print("don\'t reject H0'")
```

```
Taux de cholestérol dans a = [239, 225, 207, 209, 208, 223, 204, 208, 225, 212]
Taux de cholestérol dans b = [249, 253, 288, 287, 242, 295, 295, 288, 276, 247]
F_observé = 3.65
F_theorique = 3.18
reject H0
```

Le code ci-dessus montre deux groupes ”a” et ”b”, le premier étant le groupe recevant le médicament et ”b” le groupe qui reçoit un placebo. Ici, on aura 10 personnes dans chaque groupe (choix totalement arbitraire dans un intérêt de visualisation graphique).

A cet instant, les valeurs du taux de cholestérol de chaque individus est affiché en sortie du code: ”Taux de cholestérol dans a” et ”Taux de cholestérol dans b”.

- $F_{observé}$ est la statistique du test, elle est égale à 4,16.
- Les degrés de liberté (’dfn’ et ’dfd’) sont égaux: n - 1 = 10 - 1 = 9
- risque alpha: 5%

- La valeur critique, $F_{théorique}$ récupérée dans la table de Fisher est égale à 3,18.

Ainsi, on a $F_{observé} > F_{théorique} = 4,16 > 3,18$ et donc on rejette l'hypothèse H_0, ce qui signifie que le médicament à bien un effet sur le groupe a. Ce résultat se représente sur le graphique ci-dessous: