COMPLÉMENTS DE STATISTISTIQUES

Les statistiques servent à décrire les caractéristiques d'une population donnée, à étudier l'évolution d'une population au cours du temps ou encore à déduire des informations sur la population totale à partir d'un échantillon.

Le premier chapitre est intitulé :

1. Statistiques descritptives

1.1. Termes de base.

Definition 1.1. On appelera
- individu : unité d'observation
- population : ensembe des individus concernés par l'étude
- échantillon : sous-ensemble de la population dont les individus feront l'objet de l'étude
- variable : aspect de l'unité statistique que l'on va étudier
- modalité : valeur que peut prendre une variable

Exemples : espèce des arbres dans une forêt, âge des individus dans une population.

1.2. Les différents types de variables statistiques.

Intro : On n'étudie pas l'âge des individus d'une population et l'espèce des arbres de la même façon. On regroupe donc les variables ayant le même type de valeurs.

On distingue quatre types de variables :
- variable nominale : la variable est un nom. Ex : hêtre, chêne
- variable qualitative ordinale : les valeurs de la variable peuvent être ordonnées. Ex : peu salé, très salé
- variables quantitatives discrètes : les valeurs sont des entiers : 0,1,2,3
- variables quantitatives continues : les valeurs sont des réels ou supposés comme tels. Ex : âge, poids.

1.3. Distributions statistiques.

Notation 1.2. Nous notons N l'effectif de la population. Nous utilisons X, Y pour les variables, $X(i)$, $Y(j)$ pour les données individuelles. Les valeurs possibles de X sont notées x_i, l'effectif de la population ayant pour modalité x_i est noté n_i. Lorsque l'on distingue l'échantillon de la population, l'effectif de l'échantillon est alors noté n.

Ceci n'est valable que pour les variables qualitatives ou discrètes. Pour les variables continues nous réunissons les modalités dans des classes qui sont des intervalles de la forme $[a_{i-1}, a_i[$. Dans chaque classe on compte le nombre n_i d'individu : ce nombre est l'effectif de la classe.

Cas qualitatifs et discret :

Definition 1.3. On appelera distribution statistique des effectifs de la variable X l'ensemble des données (x_i, n_i) si X est qualitative ou discrète.

On représente la distribution statistique dans un tableau. Exemple : nombre d'enfants par couple.

Cas continu :

Definition 1.4. On appelera distribution statistique des effectifs de la variable X l'ensemble des données $([a_i, a_{i+1}[, n_i)$ si X est une variable continue.

Exemple : Age des individus dans un groupe.

On s'intéresse également aux fréquences, notées f_i pour étudier une population. Dans le cas discret, f_i est la fréquence de la modalité x_i et vaut $f_i = \frac{n_i}{N}$. Dans le cas continu, f_i est la fréquence de la classe $[a_{i-1}, a_i[$ et vaut $f_i = \frac{n_i}{N}$.

Les distributions statistiques des fréquences sont alors l'ensemble des couples (x_i, f_i) ou $([a_{i-1}, a_i[, f_i)$ suivant le cas.

Exemples ...

Formules :

$$\sum_{i=1}^{k} n_i = N,$$

$$\sum_{i=1}^{k} f_i = 1.$$

1.4. **Représentation graphique.**

1.4.1. *Variable nominale.* On représente les variables nominales par un diagramme en tuyau d'orgues ou un diagramme secteur.

Definition 1.5. Diagramme en tuyau d'orgues : les valeurs $x_1, x_2, \ldots$ sont jetées sur l'axe des abscisses. A chaque valeur x_i on associe un tuyau d'orgue de hauteur proportionnelle à l'effectif ou à la fréquence.

Diagramme secteur : l'échantillon est représenté par un disque et à chaue valeur x_i on associe un secteur angulaire d'ouverture proportionnelle à l'effectif. Plus précisément $\alpha_i = 360 \times f_i$.

1.4.2. *Variables ordinales.* On représente les variables ordinales par un diagramme en tuyau d'orgues en respectant l'ordre.

1.4.3. *Variables discrètes.* On représente les varibles discrètes par un diagramme bâton : on choisit une échelle pour l'axe des abscisses. On associe à chaque valeur x_i un bâton de hauteur proportionnelle à l'effectif.

1.4.4. *Variables continues.* On représente les variables continues par un histogramme :

on trace deux axes, l'axe horizontal représentant les valeurs de X. On y figure les extrémités des classes $a_0, a_1, \ldots$.

à chaque classe $[a_{i-1}, a_i[$ on associe un rectangle de base cette classe et dont l'aire est proportionnelle à l'effectif n_i de la classe.

En pratique la hauteur du rectangle se calcule par

$$h_i = \frac{n_i}{a_i - a_{i-1}}$$

et l'on porte ces hauteurs dans le tableau de distribution des effectifs, ou encore avec la densité de proportion $\frac{f_i}{a_i - a_{i-1}}$. Si l'on utilise les féquences, la hauteur $h_i = \frac{f_i}{a_i - a_{i-1}}$ est aussi appelée densité de proportion.

1.5. Fréquences cumulées. Fonction de répartition.

1.5.1. *Fréquences cumulées.* Lorsque les modalités sont ordonnées, on peut créer une nouvelle colonne au tableau de distribution et y porter les sommes successives des fréquences. Ces sommes sont appelées fréquences cumulées, notées F_i.

Lorsque X est ordinale ou discrète, F_i représente la proportion d'individu tels que $X \leq x_i$.

Lorsque X est continue, F_i représente la proportion d'individus tels que $X \leq a_i$.

1.5.2. *Fonction de répartition.*

Definition 1.6. Pour X une variable quantitative, on définit la quantité

$$F(x) = \text{proportion des individus j tels que } X(j) \leq x$$

F est croissante et $0 \leq F(x) \leq 1$.

1.5.3. *Calcul de F dans le cas continu.* En premier lieu $F(a_i) = F_i$. Puis sur l'intervalle $[a_i, a_{i+1}[$ on interpole et on obtient :

$$F(x) = F(a_i) + \frac{x - a_i}{a_{i+1} - a_i} \times (F(a_{i+1}) - F(a_i))$$

Exemple : tracé d'une fonction de répartition.

1.6. Caractéristiques d'une distribution.

1.6.1. *Mode.*

Definition 1.7. Le mode est la modalité ayant le plus grand effectif

Dans le cas continu, la classe modale est la classe ayant la plus grande densité de proportion

1.6.2. *Médiane.*

Definition 1.8. La médiane est la modalité qui permet de séparer en deux la population de manière égale.

Cas continu : la médiane est la valeur x telle que $F(x) = 0,5$.

Calcul pratique : on cherche la classe telle que Méd $\in [a_i, a_{i+1}[$, puis on trouve la médiane par

$$\text{Méd } = a_i + \frac{0,5 - F(a_i)}{F(a_{i+1}) - F(a_i)} \times (a_{i+1} - a_i)$$

1.6.3. *Quantiles, quartiles.* Les quantiles généralisent la notion de médiane.

Definition 1.9. Soit α un paramètre entre 0 et 1. Le quantile d'ordre α est la quantité Q_α telle que

$$F(Q_\alpha) = \alpha.$$

Calcul du quantile d'ordre α : on cherche la classe telle que $Q_\alpha \in [a_i, a_{i+1}[$, puis

$$Q_\alpha = a_i + \frac{\alpha - F(a_i)}{F(a_{i+1}) - F(a_i)} \times (a_{i+1} - a_i)$$

Definition 1.10. Les quartiles sont des quantiles pour les valeurs 0.25, 0.5, et 0.75 de α. On parle de 1er, 2ème et 3ème quartiles.

La médiane est le quantile d'ordre 0.5 et le deuxième quartile.

1.6.4. *Moyenne.*

Definition 1.11. Si X est une variable quantitative dont on possède des données individuelles $X(j)$, alors la moyenne de X est donnée par

$$\bar{x} = \frac{1}{N}(X(1) + \ldots + X(N)).$$

Si X est quantitative discrète pour laquelle on dispose de la distribution des effectifs, la formule devient

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{k} n_i x_i,$$

et avec les fréquences :

$$\bar{x} = \sum_{i=1}^{k} f_i x_i.$$

Enfin si X est continue et si l'on dispose des données regroupées en classes, on pose

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{k} n_i c_i$$

avec $c_i = \frac{a_i - a_{i-1}}{2}$ le milieu de la classe $[a_{i-1}, a_i[$. Avec les fréquences cela donne

$$\bar{x} = \sum_{i=1}^{k} f_i c_i.$$

1.6.5. *Variance, écart-type.*

Definition 1.12. Pour X variable quantitative, la variance est définie par

$$V(X) = \frac{1}{N}\sum_{i=1}^{N}(X(i) - \bar{x})^2$$

La variance peut également être calculée par la fomule :

$$V(X) = \frac{1}{N}\sum_{i=1}^{N} X_i^2 - \bar{x}^2.$$

Avec les effectifs les formules deviennent

$$V(X) = \frac{1}{N}\sum_{i=1}^{k} n_i(x_i - \bar{x})^2 = \frac{1}{N}\sum_{i=1}^{k} n_i x_i^2 - \bar{x}^2.$$

Et avec les fréquences

$$V(X) = \sum_{i=1}^{k} f_i(x_i - \bar{x})^2 = \sum_{i=1}^{k} f_i x_i^2 - \bar{x}^2.$$

Dans le cas continu, on utilise les milieux des classes c_i en lieu et place des x_i pour les formules avec les effectifs ou les fréquences.

Exemple :

$$V(X) = \frac{1}{N}\sum_{i=1}^{k} n_i(c_i - \bar{x})^2 = \frac{1}{N}\sum_{i=1}^{k} n_i c_i^2 - \bar{x}^2.$$

A partir de la variance on calcule l'écart-type :

Definition 1.13. L'écart-type est la racine carrée de la variance :

$$\sigma(X) = \sqrt{V(X}.$$

1.6.6. *Indicateurs de tendance centrale, indicateurs de dispersion.* Les indicateurs de tendance centrale sont le mode, la médiane et la moyenne et indiquent autour de quelle valeur se répartissent les valeurs de la variable.

Les indicateurs de dispersion sont l'écart-type et l'écart interquartile égal à $Q_3 - Q_1$. Ils indiquent la dispersion des valeurs autour de la tendance centrale.

2. Variables aléatoires

2.1. Variables discrètes.

Definition 2.1. Une variable X discrète est une variable aléatoire à valeurs entières. La loi d'une telle variable X est donnée par l'ensemble des $p_k = P(X = k)$.

Exemple : prenons un dé à 6 faces. Les valeurs possibles de X sont les entiers $1, \ldots, 6$, et pour chacun de ces entiers la probabilité que X ait cette valeur est

$$P(X = k) = \frac{1}{6}.$$

On introduit maintenant l'espérance d'une variable, qui correspond à la moyenne pour les échantillons :

Definition 2.2. L'espérance d'une variable discrète est

$$E(X) = \sum_k k p_k = \sum_k k \cdot P(X = k).$$

De manière plus générale on définit pour une fonction f quelconque :

$$E(f(X)) = \sum_k f(k) p_k.$$

Definition 2.3. La variance d'une variable discrète est

$$V(X) = \sum_k k^2 p_k - E(X)^2 = E[(X - E(X))^2]$$

2.2. Lois discrètes usuelles.

Definition 2.4. Soit p un paramètre entre 0 et 1. La variable X suit la loi de Bernoulli de paramètre p, notée $\mathcal{B}(p)$ si $P(X = 0) = 1 - p$ et $P(X = 1) = p$.

tt

Definition 2.5. Soit p un paramètre entre 0 et 1 et n un entier positif. La variable X suit la loi binomiale $\mathcal{B}(n, p)$ si pour tout k de $[0, n]$,

$$P(X = k) = C_N^k p^k (1 - p)^{n-k}.$$

Vérifier que la somme des $P(X = k)$ donne bien 1.

Definition 2.6. Soit λ un réel positif. La variable X suit une loi de Poisson de paramètre λ si pour tout k entier naturel,

$$P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}.$$

(1) Vérifier que la somme donne 1.
(2) Calculer $E(X)$ et $V(X)$ pour les lois usuelles.

2.3. **Variables continues.**

Definition 2.7. Une variable continue est une variable aléatoire à valeurs dans $\mathbb{R}$. Sa loi est une fonction f positive et d'intégrale 1.

La probabilité d'un événement A est alors donnée par la formule

$$P(X \in A) = \int_A f(x)dx.$$

Definition 2.8. L'espérance de X variable continue est donnée par

$$E(X) = \int_{\mathbb{R}} xf(x)dx.$$

La variance de X est donnée par

$$V(X) = \int_{\mathbb{R}} x^2 f(x)dx - E(X)^2 = E(X^2) - E(X)^2.$$

2.4. **Quelques lois continues usuelles.**

Definition 2.9. Une variable X suit la loi uniforme sur un intervalle $[a, b]$, notée $U([a, b])$, si sa densité est

$$f(x) = \frac{1}{b-a} 1_{[a,b]}(x).$$

Tracer la densité.
On définit à présent la loi exponentielle

Definition 2.10. Une variable X suit la loi exponentielle de paramètre λ, notée $\mathcal{E}(\lambda)$, si sa densité est

$$f(x) = \lambda \exp(-\lambda x) 1_{\{x \geq 0\}}.$$

Tracer la densité.

(1) Calculer l'espérance et la variance de la loi uniforme sur $[a, b]$.
(2) Donner un exemple pour la loi uniforme sur $[0, 1]$.
(3) Vérifier que nous avons bien une densité pour la loi exponentielle.
(4) Calculer l'espérance de la loi exponentielle
(5) Donner la formule

$$\int_0^\infty x^n e^{-x} dx = n!$$

2.5. **Propriétés.**

Definition 2.11. Deux variables X et Y sont indépendantes si elles n'ont aucune influence l'une sur l'autre, autrement dit si

$$P(X \in A, Y \in B) = P(X \in A) \cdot P(Y \in B)$$

pour tout événements A et B.

Les propriétés de l'espérance et de la variance sont les suivantes :
- $E(X + Y) = E(X) + E(Y)$
- $E(\lambda X) = \lambda E(X)$
- $V(\lambda X) = \lambda^2 V(X)$
- $V(X + a) = V(X)$ pour toute constante a.
- $V(X + Y) = V(X) + V(Y)$ si X et Y sont indépendantes.

2.6. **La gaussienne.**

Definition 2.12. Une variable X est normale de loi $\mathcal{N}(\mu, \sigma^2)$ si elle a pour densité

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp(\frac{-(x-\mu)^2}{2\sigma^2}).$$

La variable de loi $\mathcal{N}(0,1)$ est appelée gaussienne centrée réduite.

- $E(X) = \mu$
- $V(X) = \sigma^2$

Theorem 2.13. *Soit $X \sim \mathcal{N}(\mu, \sigma^2)$. Alors*

- $X + a \sim \mathcal{N}(\mu + a, \sigma^2)$
- $\lambda X \sim \mathcal{N}(\lambda\mu, \lambda^2\sigma^2)$.

Definition 2.14. On appelle centrée réduite de X la transformée linéaire de X ayant pour espérance 0 et pour variance 1, soit donc :
$Y = \frac{X-\mu}{\sigma}$.
Lorsque X est gaussienne, Y a pour loi $\mathcal{N}(0,1)$:
$Y = \frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$.

En additionnant deux gaussiennes indépendantes, on retombe sur une gaussienne :

Theorem 2.15. *Soit X et Y deux gaussiennes indépendantes $\mathcal{N}(\mu, \sigma_1^2)$ et $\mathcal{N}(\nu, \sigma_2^2)$. Alors $X + Y$ est une gaussienne de loi $\mathcal{N}(\mu + \nu, \sigma_1^2 + \sigma_2^2)$.*

Le théorème suivant signifie le caratère universel de la loi gaussienne :

Theorem 2.16. *(TCL) Soit $X_1, \ldots, X_n$ des variables iid de même loi que X. Notons μ l'espérance de X, et σ^2 sa variance. Pour $n \geq 30$, nous considérons que pour toute variable X :*

$$\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} \sim \mathcal{N}(0,1).$$

2.7. **Autre lois importantes.**

Definition 2.17. Si $X_1, \ldots, X_n$sont des $\mathcal{N}(0,1)$ indépendentes, alors $X_1^2 + \ldots + X_n^2$ suit la loi du chi-deux de paramètre n, notée $\chi^2(n)$.

Soit ensuite Y une $\mathcal{N}(0,1)$ indépendante de Z une $\chi^2(n)$. Alors

$$\frac{Y}{\sqrt{Z/n}}$$

suit une loi de student de paramètre n, notée $T(n)$.

3. Estimation ponctuelle

3.1. **Cadre.** On cherche à estimer un paramètre inconnu θ d'une variable aléatoire X. Une idée naturelle est d'estimer θ à l'aide d'une suite d'observations $(X_1, \ldots, X_n)$ de même loi que X.

Definition 3.1. un échantillon aléatoire ou échantillon variable est une suite $(X_1, \ldots, X_n)$ de variables aléatoires indépendantes et de même loi que X.

En pratique on dispose de données sur un seul échantillon.

Definition 3.2. un échantillon observé $(x_1, \ldots, x_n)$ est une réalisation de l'échantillon variable $(X_1, \ldots, X_n)$.

A partir des échantillons on effectue des estimations :

Definition 3.3. un estimateur est une fonction des $X_1, \ldots, X_n$.

Une estimation est la valeur numérique de l'estimateur calculée sur l'observation $(x_1, \ldots, x_n)$.

L'estimateur est dit sans biais si son espérance est égale au paramètre recherché.

3.2. **Estimation ponctuelle de la moyenne.** On considère une variable X d'espérance μ que l'on ne connait pas. On dispose d'un échantillon $(x_1, \ldots, x_n)$ de taille n qui est une réalisation de l'échantillon variable $(X_1, \ldots, X_n)$.

Definition 3.4. L'estimateur ponctuelle de la moyenne μ est

$$\bar{X} = \frac{1}{n}(X_1 + \ldots + X_n)$$

C'est la moyenne empirique.

L'estimation ponctuelle de la moyenne μ est

$$\bar{x} = \frac{1}{n}(x_1 + \ldots + x_n)$$

C'est la moyenne de l'échantillon $(x_1, \ldots, x_n)$.

Pour pouvoir discuter de la qualité de l'approximation $\bar{x}$, il faut connaître le comportement de la variable aléatoire $\bar{X}$.

Propriétés 1 :

- $E(\bar{X}) = \mu$
- $V(\bar{X}) = \frac{1}{n}V(X)$

La première ligne signifie que $\bar{X}$ est un estimateur sans biais de μ. La seconde signifie que plus n est grand plus la variable $\bar{X}$ est concentrée autour de l'espérance μ.

On peut dans certains cas avoir des informations sur la loi de $\bar{X}$.

Propriétés 2 :

- si $n \geq 30$ alors $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$, ou encore $\sqrt{n}\frac{\bar{X}-\mu}{\sigma} \sim \mathcal{N}(0,1)$.
- si X est gaussienne alors $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$, ou encore $\sqrt{n}\frac{\bar{X}-\mu}{\sigma} \sim \mathcal{N}(0,1)$.

3.3. **Estimation ponctuelle de la variance.**

Definition 3.5. L'estimation ponctuelle de la variance est

$$S^{*2} = \frac{1}{n-1}\sum_i (X_i - \bar{X})^2$$

- $E(S^{*^2}) = \sigma^2$: c'est un estimateur sans biais.
- Lorsque X est gaussienne,

$$(n-1)\frac{S*^2}{\sigma^2} \sim \chi^2(n-1).$$

- Lorsque $n \geq 30$, on peut remplacer σ par S^* dans le théorème de la limite centrale : pour X quelconque,

$$\sqrt{n}\frac{\bar{X}-\mu}{S^*} \sim \mathcal{N}(0,1).$$

Lemma 3.6. *Pour tout m nous avons*

$$\sum_{k=1}^{n}(X_k - \bar{X})^2 = \sum_{k=1}^{n}(X_k - m)^2 - n(\bar{X} - m)^2.$$

4. Estimation de la moyenne par intervalle de confiance

4.1. **Intervalle de confiance.** L'estimation ponctuelle n'est significative que si l'on connaît la probabilité d'erreur qu'elle entraîne. Pour contrôler cette probabilité on utilise la notion d'intervalle de confiance.

Definition 4.1. Soit α un paramètre proche de 0. Un intervalle de variation au niveau $1-\alpha$pour une variable Z est un intervalle $[-b,a]$ tel que $P(Z\in[-b,a])\geq 1-\alpha$. Lorsque Z est symétrique, on prendra l'intervalle de la forme $[-a,a]$.

Donner des exemples avec la gaussienne pour 80%, 95%.

Ces intervalles de variations seront utiles pour déterminer les intervalles de confiance :

Definition 4.2. Soit α un paramètre proche de 0. Un intervalle de confiance au niveau $1-\alpha$ pour le paramètre θ est un intervalle aléatoire $I(\theta)$ tel que

$$P(\theta\in I(\theta))\geq 1-\alpha.$$

Déterminations pratiques :

- soit Z une gaussienne centrée réduite. Pour donner l'intervalle de confiance au niveau $1-\alpha$, on cherche $I=[-a,a]$ tel que $P(Z>a)\leq\alpha/2$. On a bien
$$P(-a\leq Z\leq a)\geq 1-\alpha.$$
- soit $Z=\sqrt{n}\frac{\bar{X}-\mu}{S^*}$. Si l'on a un intervalle de confiance pour l'espérance de Z de type $[-b,a]$, alors

$$P(-b\leq Z\leq a)\geq 1-\alpha\Leftrightarrow P(-b\leq\sqrt{n}\frac{\bar{X}-\mu}{S^*}\leq a)\Leftrightarrow P(\bar{X}-b\cdot\frac{S^*}{\sqrt{n}}\leq\mu\leq\bar{X}+a\cdot\frac{S^*}{\sqrt{n}})$$

et alors l'intervalle de confiance pour l'espérance de X est

$$I_{1-\alpha}(\mu)=[\bar{X}-b\cdot\frac{S^*}{\sqrt{n}};\bar{X}+a\cdot\frac{S^*}{\sqrt{n}}].$$

Par la suite nous cherchons un intervalle de confiance pour μ l'espérance de X au niveau $1-\alpha$. Deux cas se présentent.

4.2. **Pour une variable quelconque n>=30.** Soit X une variable quelconque et supposons que la taille de l'échantillon n est supérieur à 30. En ce cas nous savons que

$$Z=\sqrt{n}\frac{\bar{X}-\mu}{S^*}\sim\mathcal{N}(0,1).$$

Nous cherchons a dans les tables tel que $P(Z<a)\geq 1-\alpha/2$. L'intervallle de confiance est alors

$$I_{1-\alpha}(\mu)=[\bar{X}-a\cdot\frac{S^*}{\sqrt{n}};\bar{X}+a\cdot\frac{S^*}{\sqrt{n}}]$$

4.3. **Pour une variable gaussienne n<30.** Soit X une gaussienne et n inférieur à 30. En ce cas on ne peut approximer σ par S^*. Cette fois

$$Z=\sqrt{n}\frac{\bar{X}-\mu}{S^*}\sim T(n-1).$$

Nous cherchons a dans les tables tel que $P(|Z|>a)\leq\alpha$. L'intervalle de confiance est alors

$$I_{1-\alpha}(\mu)=[\bar{X}-a\cdot\frac{S^*}{\sqrt{n}};\bar{X}+a\cdot\frac{S^*}{\sqrt{n}}]$$

5. Tests sur la moyenne

5.1. Problématique.

5.1.1. *Généralités sur les tests.* On veut comparer la moyenne inconnue μ d'une variable quantitative X à une valeur théorique donnée μ_0. On se donne donc deux hypothèses H_0 et H_1, l'hypothèse H_0 étant l'hypothèse de départ, et l'on se fixe α un paramètre entre 0 et 1.

Tester H_0 contre H_1 au niveau $1-\alpha$ consiste à définir un critère à partir d'un échantillon de telle sorte que la probabilité de rejeter H_0 alors que H_0 est vraie soit inférieure à α.

Pour ce faire on introduit une statistique Z à laquelle on associe les zones de rejet et d'acceptation. Si la statistique se trouve dans la zone de rejet nous rejettons H_0 au profit de H_1, si la statistique est dans la zone d'acceptation nous gardons H_0.

Exemple : on veut savoir au niveau 95% si la taille moyenne de la population est de 1m71 ou bien est plus grande. En ce cas $H_0 = \{\mu_0 = 1m71\}$ et $H_1 = \mu > \mu_0$. Si l'on dispose de 100 échantillons, on acceptera H_0 par exemple si $\bar{X} < 1m745$. Il faut trouver une méthode pour déterminer la zone de rejet.

5.1.2. *Tests sur la moyenne.* La statistique utilisée est

$$Z = \sqrt{n}\frac{\bar{X} - \mu_0}{S^*}.$$

L'hypothèse H_0 sera toujours $\mu = \mu_0$. On distingue trois types pour H_1 qui conduisent à différentes zones de rejets.

$H_1 = \mu > \mu_0$: alors la zone de rejet est $\{Z > a_0\}$.
$H_1 = \mu < \mu_0$: alors la zone de rejet est $\{Z < a_0\}$.
$H_1 = \mu \neq \mu_0$: alors la zone de rejet est $\{|Z| > a_0\}$.
La quantité a_0 est appelée le seuil du test.

5.2. Variable quelconque n>=30.

Sous H_0, nous avons $Z \sim \mathcal{N}(0,1)$. Si $H_1 = \mu > \mu_0$, alors Z aura tendance à être plus grande sous H_1 que sous H_0. La zone de rejet est donc de la forme $Z > a_0$. Le seuil a_0 est donné par le critère

$$P_{H0}(Z > a_0) \leq \alpha.$$

5.3. Variable gaussienne n<30.

Sous H_0, nous avons $Z \sim \mathcal{T}(n-1)$. Si $H_1 = \mu > \mu_0$, alors Z aura tendance à être plus grande sous H_1 que sous H_0. La zone de rejet est donc de la forme $Z > a_0$. Le seuil a_0 est donné par le critère

$$P_{H0}(Z > a_0) \leq \alpha.$$

5.4. Compléments sur les test.

5.4.1. *Erreurs de première et seconde espèces.*

Definition 5.1. L'erreur de première espèce est la probabilité de ne pas choisir H_0 alors que H_0est vraie. C'est donc α.

L'erreur de seconde espèce, notée β, est la probabilité de ne pas choisir H_1 alors que H_1 est vraie.

La puissance d'un test π est la probabilité de choisir H_1 lorsque H_1 est vraie. Nous avons donc $\pi = 1 - \beta$.

L'idéal serait d'avoir les plus petites erreurs possibles. Seulement α et β évoluent en sens contraire : lorsque l'on diminue α on augmente β et inversement.

5.4.2. *p-value.* Les tests sont imbriqués les uns dans les autres. Ainsi si $\alpha' \geq \alpha$ et que l'on rejette H_0 au niveau α, alors on rejette H_0 au niveau α'. Plusαest grand, plus on a de chance de rejeter H_0.

Definition 5.2. La p-value d'un test est la plus grande valeur de α telle que l'on accepte H_0.

Evaluation pratique : soit z_{obs} la valeur observée de Z. Alors la p-value est $P_{H_0}(Z > z_{\text{obs}})$ pour un test unilatéral à droite.

5.5. **Comparaisons des moyennes.** On dispose de deux échantillons $X_1, \ldots, X_n$ et $Y_1, \ldots, Y_p$ sur deux populations distinctes dont on souhaite comparer les moyennes. On suppose que ces échantillons sont gaussiens et ont mêmes variances. L'hypothèse H_0 est ici $m_X = m_Y$.

Nous utilisons la statistique

$$T = \sqrt{\frac{np(n+p-2)}{n+p}} \frac{\bar{X} - \bar{Y}}{\sqrt{\sum_{k=1}^{n}(X_k - \bar{X}) + \sum_{k=1}^{p}(Y_k - \bar{Y})^2}}$$

Sous H_0, la variable T suit la loi $\mathcal{T}(n+p-2)$.

6. Etude de la proportion

6.1. **Estimation ponctuelle d'une proportion.** On rencontre souvent le cas où une variable X peut prendre deux modaliés A et B, par exemple pour savoir si les pièces fabriquées par une machine sont défectueues ou non, ou encore pour savoir qui va être élu au second tour d'une élection (quand il n'y a plus que deux candidats). On cherche alors à estimer p la proportion d'individus ayant la modalité A dans la population totale à partir d'un échantillon.

Definition 6.1. L'estimateur ponctuel de p est la statistique

$$F = \frac{1}{n} N_A$$

où N_A st l'effectif variable de modalités A dans l'échantillon.

Propriétés :
- Cet estimateur est sans biais : $E(F) = p$.
- L'écart-type est $\sigma(F) = \sqrt{\frac{pq}{n}}$.
- Grâce au TLC, si $n \geq 30$, et si de plus $np \geq 5$ et $nq \geq 5$, alors on pourra considérer que

$$\frac{F-p}{\sqrt{\frac{pq}{n}}} \sim \mathcal{N}(0,1),$$

et également

$$\frac{F-p}{\sqrt{\frac{F(1-F)}{n}}} \sim \mathcal{N}(0,1).$$

6.2. **Intervalle de confiance pour une proportion.** Soit α un paramètre proche de 0. Pour trouver un intervalle de confiance pour la proportion p, on utilise la convergence

$$\frac{F-p}{\sqrt{\frac{F(1-F)}{n}}} \sim \mathcal{N}(0,1).$$

Au niveau α, la table de la loi $\mathcal{N}(0,1)$ nous fournit la valeur a telle que pour $Z \sim \mathcal{N}(0,1)$

$$P(|Z| \leq a) \geq 1-\alpha.$$

L'intervalle de confiance est alors donné par

$$I_{1-\alpha}(p) = \left[F - a\sqrt{\frac{F(1-F)}{n}}, F + a\sqrt{\frac{F(1-F)}{n}}\right]$$

6.3. **Test pour une proportion.** On considère l'hypothèse $H_0 = \{p = p_0\}$. Si $np_0 \geq 5$ et $n(1-p_0) \geq 5$, alors nous savons que sous H_0,

$$\frac{F-p_0}{\sqrt{\frac{p_0 q_0}{n}}} \sim \mathcal{N}(0,1).$$

La suite est similaire aux tests sur la moyenne.

7. Etude de couples de variables

7.1. Caractéristiques de distribution d'un couple.

7.1.1. *Loi jointe.* On considère sur une population donnée un couple de variable (X,Y). La cardinal du couple de modalité (x_i, y_j) est noté n_{ij}. Le cardinal de la modalité x_i est noté $n_{i.}$, celui de y_j est noté $n_{.j}$.

Definition 7.1. On appelle distirbution marginales du couple (X,Y) les distributions séparées de X et de Y. Elles sont notées $(x_i, n_{i.})$ et $(y_j, n_{.j})$ respectivement.

Formules : $N = \sum_{ij} n_{ij}$, $n_{i.} = \sum_j n_{ij}$, $n_{.j} = \sum_i n_{ij}$.
On représente toutes ces quantités dans un tableau de contingence :

7.1.2. *Distributions conditionnelles.*

Definition 7.2. la distribution conditionnelle de Y lorsque $X = x_i$ est la distribution de Y dans la sous-population vérifiant $X = x_i$. La vairable résultante est notée $Y/X = x_i$, sa loi est donnée par la $i^{\text{ème}}$ ligne.

7.1.3. *Différentes notions de fréquences.* La fréquence du couple (x_i, y_j) est $f_{ij} = \frac{n_{ij}}{N}$,
les fréquences marginales de X sont $f_{i.} = \frac{n_{i.}}{N}$,
les fréquences conditionnelles de Y lorsque $X = x_i$sont données par $f_{j/X=x_i} = \frac{n_{ij}}{n_{i.}}$.
Celles de X lorsques $Y = y_j$ sont données par $f_{i/Y=y_j} = \frac{n_{ij}}{n_{.j}}$.
Pour représenter les divverses distributions de fréquences de $Y/X = x_i$ on dresse le tableau des fréquences horizontales.

7.1.4. *Moyennes et variances.* $\bar{x} = \sum_i f_{i.} x_i$, $\bar{y} = \sum_{ij} y_j$

$$V(X) = \left(\frac{1}{N} \sum_{i=1}^{k} n_{i.} x_i^2 \right) - \bar{x}^2$$

$\bar{y}/X = x_i = \frac{1}{n_{i.}} \sum_j n_{ij} y_j = \sum_j f_j/_{X=x_i} y_j$
$V(Y/_{X=x_i}) = (\frac{1}{n_{i.}} \sum_j n_{ij} y_j^2) - (\bar{y}/_{X=x_i})^2$.

7.2. **Régression linéaire.**

7.2.1. *Covariance, corrélation.*

Definition 7.3. La covariance de X et Y est

$$cov(X,Y) = \frac{1}{N} \sum_{ij} n_{ij}(x_i - \bar{x})(y_j - \bar{y}) = \frac{1}{N} \sum_{ij} n_{ij} x_i y_j - \bar{x}\bar{y}.$$

Propriétés : $cov(\lambda X, Y) = \lambda cov(X,Y)$.
$cov(X,Y) = cov(Y,X)$.
$cov(X, X = Var(X)$

Definition 7.4. Le coefficient de corrélation est

$$r(X,Y) = \frac{cov(X,Y)}{\sigma(X)\sigma(Y)}.$$

Propriétés : $r(X,Y) \in [-1,1]$
$r(Y,X) = r(X,Y)$
$r(X,X) = 1$
$r(\lambda X, Y) = sgn(\lambda) r(X,Y)$

7.2.2. *Droite de régression linéaire.* La droite de régression linéaire est la droite qui passe au plus près du nuage de points.

la droite de régression linéaire de Y par rapport à X est donnée par

$$y - \bar{y} = \frac{cov(X,Y)}{V(X)}(x - \bar{x}).$$

Elle minimise $\sum_i n_{i.}(\bar{y}/_{X=x_i} - f(x_i))^2$.

Notons la $y = ax + b$.

Definition 7.5. Soit $\hat{Y}$ la variable de distribution $(\hat{y}_i, n_{i.})$avec $\hat{y}_i = ax_i + b$.C'est l'explication de Y par rapport à X. Cette variable est moins variable que la variable initiale Y.

Proposition 7.6. *Nous avons*

$$\frac{V(\hat{Y})}{V(Y)} = r(X,Y)^2.$$

Si $r^2 \geq 0,9$ on dit que l'ajustement linéaire est bon.

8. Tests du χ^2

8.1. **Test du χ^2 d'ajustement.** Soit X une variable aléatoire à k modalités. Pour une loi de probabilité donnée $f = \{f_1, \ldots, f_k\}$, on veut tester l'hypothèse H_0 : "X a pour loi f " contre H_1 : "X n'a pas pour loi f ". Notons n_i^X l'effectif de la modalité i dans l'échantillon, $f_i^X = \frac{n_i X}{n}$ sa fréquence et $d_i = nf_i$ l'effectif théorique de la modalité i dans l'échantillon.

Le test d'ajustement est basé sur la distance

$$D = \sum_{i=1}^{k} \frac{(d_i - n_i^X)^2}{d_i}$$

Theorem 8.1. *Sous H_0, si les d_i sont tous ≥ 5, on admet que $D \sim \chi^2(k-1)$.*

8.2. **Test du χ^2 d'homogénéïté.** Le test d'homogénéïté est une version améliorée du test d'ajustement. On cherche à savoir si deux variables X et Y ont la même loi.

Soit donc deux variables X et Y ayant les mêmes modalités, au nombre de k. On veut tester l'hypothèse H_0 : "X et Y ont la même loi " contre H_1 : "X et Y n'ont pas la même loi ". Si les deux variables avaient même loi, les fréquences respectives seraient proches de celles trouvées dans l'échantillon total. Posons donc

$$\hat{f}_i = \frac{n_i^X + n_i^Y}{n+p}.$$

Les effectifs théoriques dans chaque échantillon sont alors $d_i^X = n \times \hat{f}_i$ et $d_i^Y = p \times \hat{f}_i$.

Posons

$$D^X = \sum_{i=1}^{k} \frac{(d_i^X - n_i^X)^2}{d_i^X}$$

et

$$D^Y = \sum_{i=1}^{k} \frac{(d_i^Y - n_i^Y)^2}{d_i^Y}.$$

Theorem 8.2. *Sous H_0, si tous les d_i^X et d_i^Y sont ≥ 5, alors on admet que $D = D^X + D^Y \sim \chi^2(k-1)$.*