III - INTRODUCTION AUX TESTS STATISTIQUES

J-P. Croisille

Université de Lorraine

UEL - Année 2012/2013

1-PRINCIPE DES TESTS D'HYPOTHESE

HYPOTHESE NEUTRE ET HYPOTHESE ALTERNATIVE:

Une **hypothèse** est une affirmation que quelque chose à propos d'une population est VRAI. Le problème est que l'on ne dispose que d'un échantillon de cette population. On veut simplement au vu de cet échantillon *se faire une idée* de la vérité de cette affirmation. Un test statistique se présente sous la forme de 2 hypothèses en concurrence:

- l'*hypothèse neutre*, notée H_0
- l'*hypothèse alternative*, notée H_a.

Le problème que l'on souhaite résoudre avec un **test d'hypothèse** est de décider de façon rationnelle au vu d'un échantillon si l'hypothèse neutre doit être rejetée en faveur de l'hypothèse alternative.

CHOIX DE L'HYPOTHESE NEUTRE ET DE L'HYPOTHESE ALTERNATIVE: EXEMPLE 1

D'après des données des libraires, le prix moyen des livres d'histoire vendus en France a été de 35.27 euros en 2000. On veut savoir si le prix moyen des livres d'histoire vendus en France en 2008 est supérieur à 35.27 euros.

1. Détermination de l'hypothèse neutre H_0
   L'hypothèse neutre est:
   (H_0) :" Le prix moyen des livres d'histoire vendus en France en 2008 est $\mu = 35.27$ euros".
2. Détermination de l'hypothèse alternative H_a
   L'hypothèse alternative choisie ici est:
   (H_a) :" Le prix moyen des livres d'histoire vendus en France en 2008 est $\mu > 35.27$ euros". On dit que l'hypothèse alternative est *unilatérale à droite* car le signe $>$ apparaît dans l'hypothèse alternative.

CHOIX DE L'HYPOTHESE NEUTRE ET DE L'HYPOTHESE ALTERNATIVE: EXEMPLE 2

La prise de calcium journalière recommandée est de 1000 mg par jour. On souhaite effectuer un test statistique sur un échantillon de personnes pour savoir si le fait d'avoir un revenu au dessous du seuil de pauvreté a une influence sur ce nombre.
On appelle μ la moyenne de la prise quotidienne de calcium de tous les individus ayant un revenu au dessous de seuil de pauvreté.

1. **Détermination de l'hypothèse neutre** H_0
   L'hypothèse neutre est:
   (H_0) :" La prise de calcium journalière moyenne pour tous les adultes vivant au dessous du seuil de pauvreté est $\mu = 1000$ mg".
2. L'hypothèse alternative **choisie** est:
   (H_a) :" La prise de calcium journalière moyenne pour tous les adultes vivant au dessous du seuil de pauvreté est $\mu < 1000$ mg". On dit que l'hypothèse alternative est *unilatérale à gauche* car le signe $<$ apparaît dans l'hypothèse alternative.

LOGIQUE DES TESTS D'HYPOTHESE

- Prendre un échantillon d'une population.
- Si les données de l'échantillon ne sont pas en contradiction avec l'hypothèse neutre, alors l'hypothèse neutre n'est pas rejetée.
- Si les données de l'échantillon sont en contradiction avec l'hypothèse neutre, alors on décide de rejeter l'hypothèse neutre et on conclut que l'hypothèse alternative est vraie.
- Analogie juridique: on se demande si l'échantillon fournit suuffisamment d'évidence pour rejeter l'hypothèse neutre au profit de l'hypothèse alternative.

TERMINOLOGIE UTILISEE DANS LES TESTS D'HYPOTHESE

- **Test statistique:** La *statistique* utilisée comme base de décision pour le rejet de l'hypothèse neutre. Ici *statistique*= quantitée évaluée dépendant de l'échantillon.
- **Région de rejet:** L'ensemble des valeurs du test statistique qui conduit au rejet de l'hypothèse neutre.
- **Région de non rejet de l'hypothèse neutre:** L'ensemble des valeurs du test statistique qui conduisent au non-rejet de l'hypothèse neutre.
- **Valeur critique:** Les valeurs du test statistique qui séparent la région de rejet et la région de non rejet de l'hypothèse neutre.

ERREUR DE TYPE I

On effectue un test sur **UN SEUL ECHANTILLON** d'une population pour donner un diagnostic à propos d'une hypothèse sur la totalité d'une population. Il est donc possible que l'on prenne la mauvaise décision si l'échantillon dont on dispose ne représente pas correctement la population.

Exemple du prix des livres d'histoire

Supposons qu'en réalité les livres d'histoire n'aient pas augmenté en moyenne entre 2000 et 2008. C'est-à-dire que l'hypothèse neutre soit VRAIE. L'échantillon dont on dispose peut avoir un prix moyen supérieur au prix moyen de l'année 2000 simplement par hasard. Dans ce cas, l'évaluation du test statistique va nous conduire à une **erreur de type I** (ou *erreur de première espèce*): on rejette à tort l'hypothèse neutre.

ERREUR DE TYPE II

Exemple de la prise journalière de calcium

Supposons qu'en réalité la prise journalière moyenne de calcium dans la population vivant au-dessous du seuil de pauvreté soit $\mu < 1000$mg. C'est-à-dire que l'hypothèse alternative soit en fait VRAIE. L'échantillon de population que l'on observe peut avoir (pour des raisons géographiques par exemple) une moyenne de prise journalière de calcium très voisine de 1000 mg. On ne pourra donc pas faire le diagnostic correct, c'est-à-dire $\mu < 1000$mg. On commet dans ce cas une **erreur de type II** (ou erreur de deuxième espèce).

TERMINOLOGIE UTILISEE DANS LES TESTS D'HYPOTHESE (2)

- **Erreur de TYPE I** : Rejeter à tort l'hypothèse neutre alors qu'elle est vraie.
- **Erreur de TYPE II** : Ne pas rejeter l'hypothèse neutre alors qu'elle est fausse.

TERMINOLOGIE UTILISEE DANS LES TESTS D'HYPOTHESE (4)

- **Erreur de TYPE I**
  La probabilité de faire une erreur de TYPE 1, c'est-a-dire de rejeter l'hypothèse neutre lorsqu'elle est vraie s'appelle le *niveau de risque (ou niveau de signification)* du test. On la note α. **C'est l'utilisateur qui choisit le niveau de risque du test qu'il effectue**.

$$\alpha = \text{niveau de risque du test.} \tag{1}$$

- **Erreur de TYPE II**
  MAIS, pour un échantillon donné, plus on choisit un niveau faible pour α, c'est-à-dire de faire une erreur de TYPE I, plus grande sera la probabilité β de ne pas rejeter une hypothèse neutre fausse, c'est-à-dire de faire une erreur de TYPE II.

TERMINOLOGIE UTILISEE DANS LES TESTS D'HYPOTHESE (5)

Supposons qu'un test d'hypothèse soit effectué avec un niveau de risque α faible. Alors

- Si l'hypothèse neutre est rejetée, on conclut que l'hypothèse alternative est vraie.
- Si l'hypothèse neutre n'est pas rejetée, on conclut que les données ne fournissent pas suffisamment d'évidence pour rejeter l'hypothèse neutre.

2- TEST DE COMPARAISON A LA MOYENNE: EXEMPLE

Exemple : Contrôle qualité (1)

Une entreprise produit des paquets de friandises de 454 grammes. En fait le poids des paquets varie d'un paquet à l'autre. Le programme qualité de l'entreprise conduit à effectuer péridiquement un test statistique sur les hypothèses suivantes:

- Hypothèse H_0: La machine d'emballage fonctionne correctement, ce qui signifie que le **poids moyen des paquets produits est** 454**g**. C'est la *situation normale*.

$$(H_0) \quad \mu = 454g \tag{2}$$

- Hypothèse H_a: La machine d'emballage ne fonctionne pas correctement, ce qui signifie que le poids moyen des paquets est différent de 454g.

$$(H_0) \quad \mu \neq 454g \tag{3}$$

L'hypothèse alternative est bilatérale car on a un signe $\neq$ dans cette hypothèse.

Exemple: Contrôle qualité (2)

Pour réaliser le test, on décide de se baser sur un échantillon de 25 paquets tirés au hasard. On obtient les poids suivants:

465	456	438	454	447
449	442	449	446	447
468	433	454	463	450
446	447	456	452	444
447	456	456	435	450

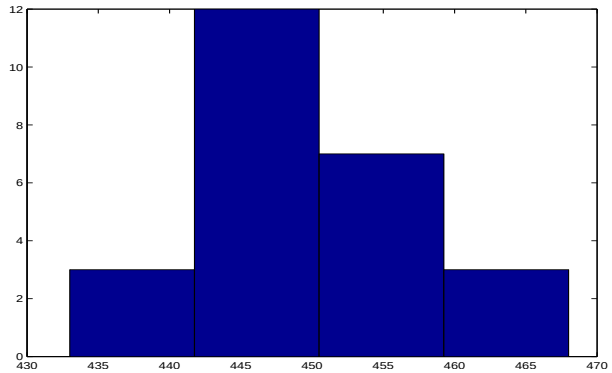

Figure: Histogramme 1

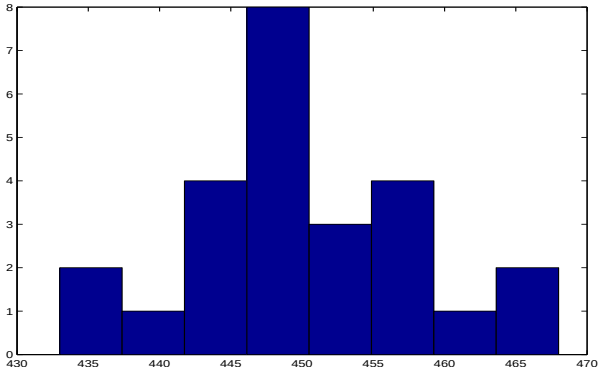

Figure: Histogramme 2

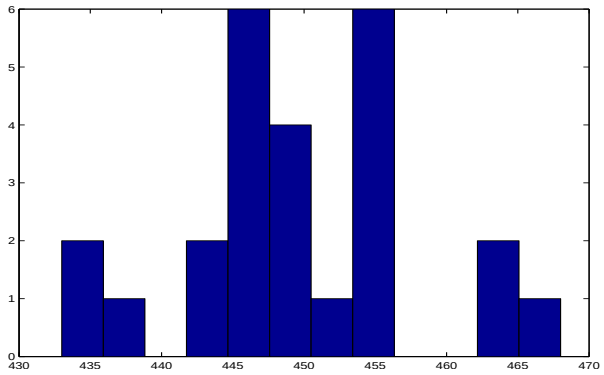

Figure: Histogramme 3

Exemple: Contrôle qualité (3)

On suppose que la distribution des poids des paquets est normale $N(\mu, \sigma)$, avec une moyenne μ que l'on ne connaît pas. Par contre, une étude préliminaire a permis de dire que l'écrat-type est $\sigma = 7.8$g.
La question est :
Est-ce que l'échantillon des 25 paquets tiré au hasard, fournit suffisamment d'évidence pour dire que la machine d'emballage ne fonctionne pas correctement ?.
C'est-à-dire:
L'Hypothèse (H_0) est REJETEE au profit de l'hypothèse (H_a).

Exemple: Contrôle qualité:

METHODE GENERALE POUR LA REALISATION D'UN TEST DE COMPARAISON A LA MOYENNE

(a) Formuler l'hypothèse neutre et l'hypothèse alternative.

(b) Discuter la logique du test.

(c) Identifier la distribution de la variable $\bar{X}$, c'est-à-dire la distribution de la moyenne empirique d'un échantillon de 25 paquets.

(d) Obtenir un critère pour décider si l'on rejette l'hypothèse neutre ou si on la conserve.

(e) Appliquer le critère (d) à l'échantillon particulier dont on dispose et formuler la conclusion du test.

Exemple: Contrôle qualité (5)

Reprenons ces étapes sur notre exemple:

(a) *Formuler l'hypothèse neutre et l'hypothèse alternative.*

On note μ la moyenne réelle des paquets. Les hypothèses choisies sont:

$$(H_0) : \mu = 454g\text{: la machine fonctionne correctement} \quad (4)$$

$$(H_a) : \mu \neq 454g\text{: la machine ne fonctionne pas correctement} \quad (5)$$

Exemple: Contrôle qualité (6)

(b) *Logique du test d'hypothèse.*

Si la moyenne $\mu = 454g$, alors la moyenne de l'échantillon va être *à peu près* égale à $454g$. Autrement dit, si la moyenne de l'échantillon diffère *trop* de 454g, on va être incliné à penser qu'il faut rejeter le fait que la moyenne de tous les paquets est de 454g.

Exemple: Contrôle qualité (7)

(c) *Identifier la distribution de la variable $\bar{X}$.*

On désigne par $\bar{X}$ la variable aléatoire correspondant à la moyenne empirique d'un échantillon tiré au hasard. On suppose ici que la variable X= poids d'un paquet tiré au hasard suit une loi normale, de moyenne μ (INCONNUE) et de variance σ_X.

Alors on sait que la variable $\bar{X}$=moyenne d'un échantillon suit AUSSI une loi normale de même moyenne μ et d'écart-type $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$.

Dans notre exemple, on suppose que la variable X est normalement distribuée. On suppose également que l'écart-type est $\sigma_X = 7.8$ g. Ces deux informations sont supposéees données par une analyse préliminaire. On en déduit:

- $\mu_{\bar{X}} = \mu$
- $\mu_{\bar{X}} = \frac{7.8}{\sqrt{25}} = 1.56$
- $\bar{X}$ est distribuée normalement.

Exemple: Contrôle qualité (8)

Méthode pour vérifier visuellement de façon préliminaire que l'échantillon est distrinué selon une loi normale.

On fait une représentation graphique de l'échantillon avec un diagramme de comparaison à une loi normale (*normal plot*).

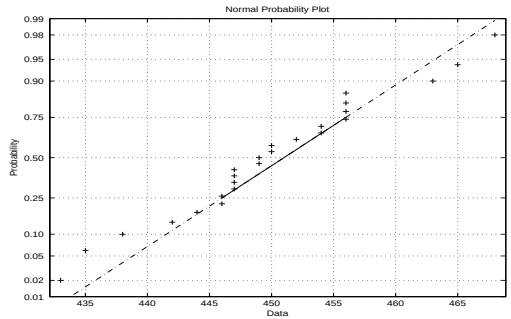

Figure: Normal plot

Exemple: Contrôle qualité (9)

On sait que si une variable X suit une distribution normale de moyenne μ et de variance σ, alors on a les probabilités suivantes

- $\mathcal{P}[\mu - \sigma, \mu + \sigma] \simeq 68.27$. C'est-à-dire, 68.27% de chances d'observer un tirage au hasard de X dans l'intervalle $[\mu - \sigma, \mu + \sigma]$.
- $\mathcal{P}[\mu - 2\sigma, \mu + 2\sigma] \simeq 95.45$. C'est-à-dire, 95.45% de chances d'observer un tirage au hasard de X dans l'intervalle $[\mu - 2\sigma, \mu + 2\sigma]$.
- $\mathcal{P}[\mu - 3\sigma, \mu + 3\sigma] \simeq 99.73$. C'est-à-dire, 99.73% de chances d'observer un tirage au hasard de X dans l'intervalle $[\mu - 3\sigma, \mu + 3\sigma]$.

Exemple: Contrôle qualité (10)

On applique ceci **à la variable** $\bar{X}$.

On en déduit que 95.44% de tous les échantillons de 25 paquets ont un poids moyen dans l'intervalle

$$[\mu - 2 \times 1.56, \mu + 2 \times 1.56] = [\mu - 3.12, \mu + 3.12] \tag{6}$$

Ceci peut s'exprimer aussi en disant que seulement $4.56\% = 100\% - 95.44\%$ de tous les échantillons de 25 sacs ont un poids moyen qui n'est pas dans l'intervalle $[\mu - 3.12, \mu - 3.12]$.

Exemple: Contrôle qualité (11)

Donc, si la moyenne de notre échantillon particulier (tiré au hasard) **n'est pas dans l'intervalle** $[454 - 3.12, 454 + 3.12]$ on a un indice très fort que la moyenne μ n'est pas 454g. Pourquoi ? Car si on avait $\mu = 454$ g, **observer un tel échantillon n'arriverait que pour** 4.54% **des échantillons de 25 sacs tirés au hasard.** Sur la base de ce raisonnement, on décide d'adopter le critère suivant pour tester notre hypothèse neutre (H_0):

Si le poids moyen de notre échantillon est en dehors de l'intervalle $[454 - 3.12, 454 + 3.12]$, alors on rejette l'hypothèse neutre et on dit que la machine d'emballage ne fonctionne pas correctement **au risque 4.54 %**. Sinon, on dit qu'on ne rejette pas l'hypothèse neutre.

Exemple: Contrôle qualité (12)

Ici, on a que le poids moyen de l'échantillon est

$$m = 450g \tag{7}$$

Pour tester l'identité $\mu = 454g$, on évalue la quantité z par

$$z = \frac{m - 454}{1.56} = \frac{450 - 454}{1.56} = -2.56 \tag{8}$$

Ceci signifie que la moyenne empirique de l'échantillon est à 2.56 fois la variance au-dessous de l'hypothèse $\mu = 454g$ (c-a-d., l'hypothèse neutre). Donc **on rejette l'hypothèse neutre avec "un risque 4.54%"**. En pratique, on dit que la machine d'emballage ne fonctionne pas correctement " au risque 4.54 %".

Exemple: Contrôle qualité (13)

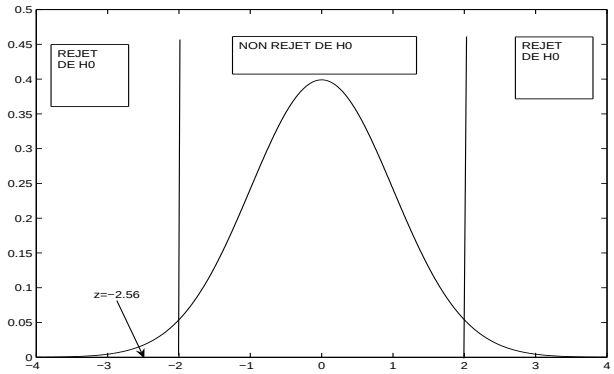

Figure: Graphe montrant la position de z par rapport aux régions de rejet et de non-rejet

3- DISTRIBUTION D'ECHANTILLONAGE

Distribution de la moyenne (1)

On considère un échantillon de taille n corespondant à une variable aléatoire X inconnue représentant des tirages dans une population.

- Caractéristiques de l'échantillon

$$\text{moyenne empirique } m = \frac{1}{n}\sum x_i, \quad \text{var. empirique } s^2 = \frac{1}{n-1}\sum (x_i - m)^2 \tag{9}$$

- Caractéritiques de la v.a.: sa fonction de distribution. En particulier sa moyenne, sa variance.

$$\mu = \int x f(x) dx, \quad \sigma^2 = \int (x-\mu)^2 f(x) dx, \tag{10}$$

Distribution de la moyenne (2)

Question: quelle est la validité de l'approximation **moyenne VRAIE de X $\simeq$ moyenne empirique de l'échantillon**.
Pour le savoir on effectue une nouvelle démarche probabiliste, celle de considérer **tous les choix possibles d'echantillons de n individus**.

- On considère sur cette population la variable aléatoire $\bar{X}$ définie par

$$\bar{X} = \sum_{i=1}^{n} X_i \tag{11}$$

  où les X_i sont n variables aléatoires de même loi que X. On appelle X la v.a. **parente**.
- On identifie ensuite la loi de $\bar{X}$ en fonction de celle de X.
  La variable aléatoire $\bar{X}$ est la variable aléatoire d'échantillonage de la moyenne (sur la population de tous les echantillons à n individus).

Distribution de la moyenne (3)

On a les résultats suivants:

- La moyenne de $\bar{X}$ est identique à la moyenne de X.

$$\mu_{\bar{X}} = \mu_X \tag{12}$$

- L'écart-type de $\bar{X}$ est

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \tag{13}$$

Autrement dit,

- la moyenne de la valeur m **sur tous les échantillons possibles de n individus** est la même que celle de la v.a. parente.
- l'écart-type tend vers 0 à la vitesse $\frac{1}{\sqrt{n}}$.

Distribution de la moyenne (4)

Exemple: Si la v.a. parente X a une loi $N(\mu, \sigma)$, alors la v.a. $\bar{X}$ a une loi $N(\mu, \frac{\sigma}{\sqrt{n}})$.

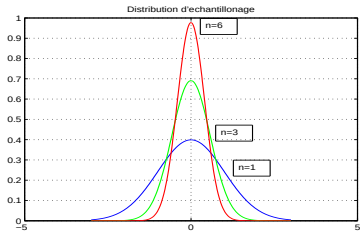

Figure: Distribution d'échantillonage d'une gaussienne: la variance diminue lorsque la taille de l'échantillon augmente.

Distribution de la moyenne (5)

Dans le cas où X **n'est pas distribuée selon une loi normale**, on a cependant le théorème de la "limite centrale".

Théorème

Si X est une v.a. de moyenne μ et de variance σ^2, et $X_1, X_2, ...X_n$ sont des v.a. de même loi que X, alors la v.a. $\bar{X}$,

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + ...X_n) \tag{14}$$

est telle que la variable aléatoire

$$\frac{\bar{X} - \mu}{\sigma\sqrt{n}} \tag{15}$$

converge en loi vers une loi normale centrée réduite $N(0, 1)$

4- LE Z-TEST DE COMPARAISON A LA MOYENNE

Méthode générale pour le Z-test

BUT: réaliser un test d'hypothèse sur la moyenne d'une population μ lorsque la variance est connue.

PRINCIPE: On sait par l'étude de la distribution d'échantillonage de la moyenne empirique, que si $\mu = \mu_0$, alors la quantité

$$z = \frac{m - \mu_0}{\sigma/\sqrt{n}} \tag{16}$$

suit une loi proche d'une loi normale $N(0, 1)$.

DONC SI L'HYPOTHESE $\mu = \mu_0$ EST VRAIE, ALORS ON PEUT OBSERVER SES CONSEQUENCES SUR L'ECHANTILLON EN OBSERVANT SA MOYENNE ET SA VARIANCE EMPIRIQUE

Méthode générale pour le Z-test

HYPOTHESES:

- L'échantillon est supposé tiré au hasard.
- La VRAIE distribution est supposée normale.
- L'écart-type σ est supposé connu, (par une étude préliminaire).

1. L'hypothèse neutre (H_0) est $\mu = \mu_0$ et l'hypothèse alternative est

$$(H_a)\ \ \mu \neq \mu_0 \quad OU(H_a)\ \ \mu < \mu_0 \quad OU(H_a)\ \ \mu > \mu_0 \tag{17}$$

2. Décider la valeur du risque α.
3. Calculer la valeur du test statistique z

$$z = \frac{m - \mu_0}{\sigma/\sqrt{n}} \tag{18}$$

4. Chercher dans la table de la **loi normale** les valeurs $\pm z_{\alpha/2}$ OU $-z_\alpha$ OU z_α.
5. Si les valeurs tombent dans la région de rejet alors rejeter (H_0), sinon, ne pas rejeter H_0.
6. Interprêter les résultats du test.

Quand utilise-t-on le Z-test? (1)

- Pour de petits échantillons (ce qui signifie en pratique $n \leq 15$), le Z-test ne doit être utilisé seulement si la variable à étudier est normale ou très proche de l'être. On doit donc faire une vérification préliminaire des données (avec un normal plot par exemple), pour voir si on peut considérer qu'il en est ainsi.
- Pour des échantillons de taille modérée, (en pratique $15 \leq n \leq 30$, le $Z-$ test peut être utilisé, sauf si l'échantillon contient des valeurs aberrantes ou si la distribution à étudier est loin d'être normale.

Quand utilise-t-on le Z-test? (2)

Pour des échantillons de grande taille ($n \geq 30$), on peut utiliser le Z-test. Cependant, si il y a des mesures qui semblent aberrantes, il est recommandé d'appliquer le test avec et sans les mesures aberrantes pour voir l'effet qu'elles ont.

5- LE T-TEST DE COMPARAISON A LA MOYENNE

Principe général pour le T-test

Si X est une v.a. de loi $N(\mu, \sigma)$. On considère X_1, X_2, ...X_n n v.a. indépendantes de même loi que X. On en déduit les deux v.a.

- La v.a. $\bar{X}$ déduite de la moyenne empirique,

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{19}$$

- La v.a. S^2 déduite de la variance empirique,

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 \tag{20}$$

Alors la v.a. T définie par

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n-1}} \tag{21}$$

suit une loi de Student a $n-1$ degrés de liberté. **La loi T_{n-1} est indépendante de la variance σ !**. On note T_{n-1} cette loi.

Distribution de Student

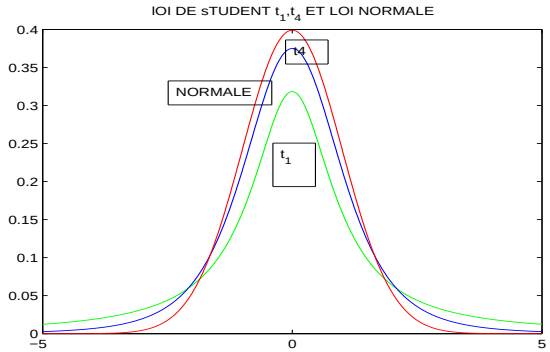

Figure: Distribution de Student et loi normale

Principe général pour le T-test (2)

Cette propriété permet de réaliser un test d'hypothèse sur la moyenne d'une population μ lorsque **la variance est inconnue**. Pour un échantillon donné, on applique la même logique de test que pour le $Z-$ test, c'est-à-dire quand on connaît σ, mais on utilise la distribution T , (loi de Student) avec $n-1$ degrés de liberté, au lieu de la loi normale centrée réduite Z.

Exemple:

Si l'hypothèse neutre est

$$(H_0) \;\; \mu = \mu_0 \tag{22}$$

on va avoir à évaluer

$$t = \frac{m - \mu_0}{s/\sqrt{n}} \tag{23}$$

On utilisera la table de Student (ligne $n-1$).

LE T-TEST: PROCEDURE GENERALE (1)

BUT: réaliser un test d'hypothèse sur la moyenne d'une population μ lorsque la variance est inconnue.

HYPOTHESES:

- L'échantillon est supposé tiré au hasard.
- La VRAIE distribution est supposée normale OU BIEN l'échantillon est grand.
- La variance σ est inconnue.

LE T-TEST: PROCEDURE GENERALE (2)

1. L'hypothèse neutre (H_0) est $\mu = \mu_0$ et l'hypothèse alternative est

$$(H_a)\ \ \mu \neq \mu_0 \quad OU \quad (H_a)\ \ \mu < \mu_0 \quad OU \quad (H_a)\ \ \mu > \mu_0 \tag{24}$$

2. Décider la valeur de α.
3. Calculer la moyenne empirique et l'écart-type empirique de l'échantillon, puis calculer la valeur du test statistique t

$$t_0 = \frac{m - \mu_0}{s/\sqrt{n}} \tag{25}$$

4. Chercher dans la table de la **loi de Student** T_{n-1} les valeurs $\pm t_{\alpha/2}$ OU $-t_\alpha$ OU t_α.
5. Si les valeurs tombent dans la région de rejet alors rejeter (H_0), sinon, ne pas rejeter H_0.
6. Interprêter les résultats du test.

EXEMPLE

On étudie statistiquement l'acidité moyenne des lacs de haute montagne des Alpes. Est-il plausible que l'acidité moyenne de ces lacs ait augmenté récemment ? Pour le savoir on mesure l'acidité d'un échantillon de 15 lacs choisis correctement. On obtient les mesures suivantes de pH.

7.2	7.3	6.1	6.9	6.6
7.3	6.3	5.5	6.3	6.5
5.7	6.9	6.7	7.9	5.8

ETAPE PRELIMINAIRE: NORMPLOT

On commence par faire une représentation des données de type *boxplot* et une autre de type *normplot*.

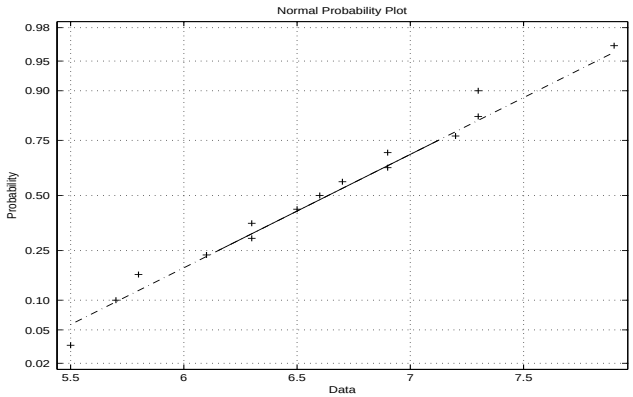

Figure: Normplot - Acidité des lacs des Alpes

ETAPE 1: FORMULATION DES HYPOTHESES

Formuler l'hypothèse neutre et l'hypothèse alternative. On note μ la moyenne réelle du pH de tous les lacs de haute montagne des Alpes. On cherche une réponse à l'alternative suivante:

$$(H_0) : \mu = 6\text{: en moyenne, les lacs sont acides} \tag{26}$$

$$(H_a) : \mu > 6\text{: en moyenne, les lacs sont non acides} \tag{27}$$

ETAPE 2: CHOIX D'UN NIVEAU DE RISQUE

On choisit un niveau de risque (ou de signification) α. On prend ici $\alpha = 0.05$.

ETAPE 3: CALCUL DU TEST STATISTIQUE T

On calcule la moyenne et l'écart-type empirique de l'échantillon. On obtient:

$$m = 6.6, \quad s = 0.672 \tag{28}$$

On obtient donc pour valeur de t

$$t = \frac{m - \mu_0}{s/\sqrt{n}} = \frac{6.6 - 6}{0.672/\sqrt{15}} = 3.458 \tag{29}$$

ETAPE 4: RECHERCHE DU T CRITIQUE SUR LA TABLE DE STUDENT

On regarde sur la table de la distribution de Student T_{14} car $n - 1 = 15 - 1 = 14$ la valeur critique $t_{0.05}$. On consate que c'est $t = 1.761$.

ETAPE 5: DIAGNOSTIC

Si le résultat du test tombe dans la zone de rejet, on rejette l'hypothèse neutre H_0, sinon on ne rejette pas H_0. Ici, on a

$$t = 3.458 > 1.761 \tag{30}$$

Donc on rejette H_0 avec un niveau de rique de 5%.

ETAPE 6: INTERPETATION

Avec un niveau de risque de 5%, les données fournissent suffisamment d'évidence pour dire que, en moyenne, les lacs des Alpes ne sont pas acides.

6- LES TESTS du χ^2

La distribution du χ^2

Une variable aléatoire Y suit une loi du χ^2 à n degrés de libertés si

$$Y = X_1^2 + X_2^2 + \cdots + X_n \tag{31}$$

où les X_i suivent des lois normales $N(0,1)$. On note $Y \sim \chi_n^2$.

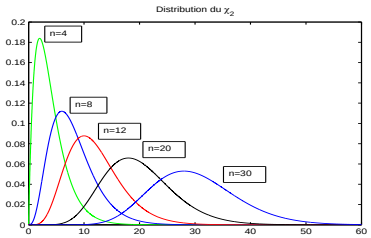

Figure: Distribution du χ_2

Propriétés de la distribution χ^2

- Moyenne: $E(\chi^2_n) = n$.
- Variance: $\sigma^2(\chi^2_n) = 2n$.
- Densité sur $\mathbb{R}^+$:

$$g(x) = \frac{1}{2^{n/2}\Gamma(\frac{n}{2})} \exp(-\frac{1}{2}x)x^{\frac{n}{2}-1} \tag{32}$$

Tests statistique et χ^2

On utilise la distribution du χ^2 dans le contexte suivant. On considère une variable aléatoire X. On considère ensuite k classes (ou k évenements) de probabilité $p_1, ..., p_k$. On a ensuite un n échantillon donnant des effectifs N_1,...N_k pour chaque classe, avec $n = N_1 + N_2 + \cdots N_k$. On a donc:

- "effectif observé dans la classe i"= N_i
- "effectif théorique dans la classe i"= $E_i = np_i$

On mesure la "distance" entre **effectif observé** et **effectif théorique** par la variable aléatoire

$$D^2 = \sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i} \tag{33}$$

On a le résultat de probablités suivant:

Proposition

La loi de variable aléatoire D^2 se comporte comme une loi χ^2_{k-1} lorque la taille de l'échantillon est grande.

LE TEST DU χ^2: PROCEDURE GENERALE (1)

BUT: réaliser un test d'hypothèse sur la conformité d'un échantillon à une répartition théorique en classes.

HYPOTHESES:

- les valeurs théoriques (attendues) sont $E_i = np_i \geq 5$.
- Au plus 20% des nombres $E_i \leq 5$.
- L' échantillonage est simple.

ETAPE 1: FORMULATION DES HYPOTHESES

Les hypothèses sont

- (H_0): la v.a. a la distribution théorique.
- (H_a): la v.a. n'a pas la distribution théorique.

ETAPE 2: CALCUL DES FREQUENCES THEORIQUES

On calcule les fréquences (les effectifs) théoriques attendues par

$$E_i = np_i \tag{34}$$

où p_i et la fréquence relative théorique.

ETAPE 3: VERIFICATION DES HYPOTHESES

Vérifier que les fréquences attendues E_i satisfont les hypothèses 1 et 2. Si ce n'est pas le cas, ne pas utiliser ce test.

ETAPE 4: CHOIX D'UN NIVEAU DE RISQUE

Décider le niveau de risque (ou de signification) α.

ETAPE 5: EVALUER LE χ^2

Evaluer le χ^2 de l'échantillon par

$$\chi_0^2 = \sum_{i=1}^{k} \frac{(N_i - E_i)^2}{E_i} \tag{35}$$

ETAPE 6: IDENTIFICATION DE LA VALEUR CRITIQUE

Identifierla valeur critique χ_α^2 dans une table du χ^2 à $k-1$ degrés de liberté.

ETAPE 7: ANALYSE DES RESULTATS

Si la valeur de χ_0^2 est dans la région de rejet, rejeter (H_0), sinon ne pas rejeter (H_0).

ETAPE 8: INTERPRETATION DES RESULTATS