Démarche Statistique 1

Introduction aux tests

Pierre Neuvial, http://stat.genopole.cnrs.fr/members/pneuvial/demstat
Evry, M1 SGO, automne 2014

La démarche de test

Problématique

Cadre

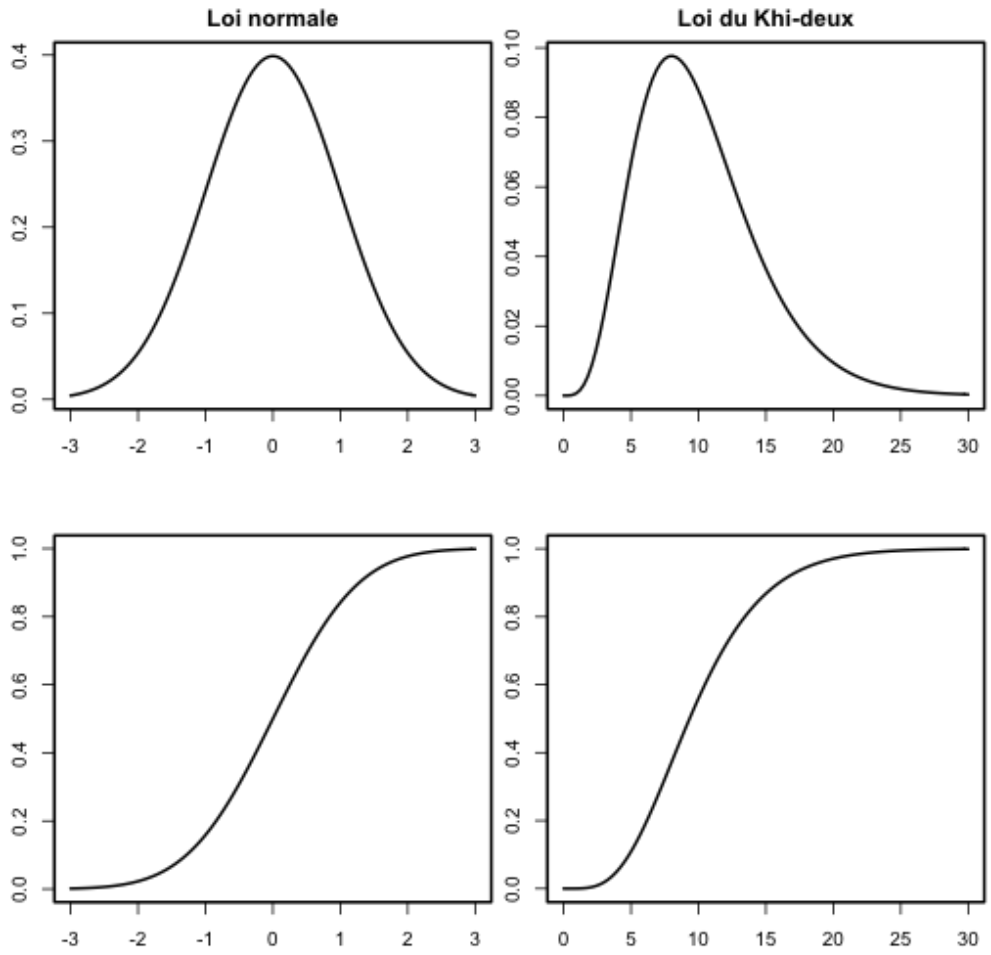
Statistique **inférentielle**

Objectif

Répondre à une question concernant un paramètre d'une population au vu d'un échantillon tiré dans cette population

Principe général

Si la distribution d'une variable aléatoire est connue, on sait dire si une observation de cette variable est exceptionnelle ou non



Exemples

Abondance en GC

Une séquence d'ADN est-elle significativement enrichie en cytosines et guanine par rapport à un modèle simple où chaque nucléotide apparait avec la même fréquence ?

Rendement d'une nouvelle variété de pommes

On veut comparer la production d'une nouvelle variété de pommes à l'ancienne variété, en termes de volume de production par arbre (en kg) et de taille des fruits (diamètre en cm). On se demande en particulier:

- la nouvelle variété est-elle plus productive que l'ancienne ?
- le diamètre des fruits est-il plus homogène pour la nouvelle variété que pour l'ancienne ?

Définition

Un test d'hypothèse (paramétrique) est constitué de 4 éléments:

1. des **données**: $x_1, x_2, \dots x_n$, réalisations de n variables aléatoires $X_1, X_2, \dots X_n$
2. un **modèle statistique**: la loi de probabilité de $X_1, X_2, \dots X_n$, dépendant d'un ou plusieurs paramètres θ
3. une **hypothèse** concernant θ, appelée *hypothèse nulle* et notée H_0
4. une **règle de décision**, composée d'
   - une *statistique de test* T fonction de $X_1, X_2, \dots X_n$
   - une *région de rejet* $\mathcal{R}$ contenant des valeurs de T improbables sous H_0
     La règle de décision est: "rejeter H_0 si $T \in \mathcal{R}$"

Remarques

- La décision est aléatoire: elle dépend de l'échantillon
- Le choix de la règle de décision sera dicté par le risque d'erreur que l'on s'autorise

Exemple: abondance en GC

1. données: séquence d'ADN de longueur n. Chaque x_i vaut 1 si la base i est G ou C, 0 sinon.
2. modèle: X_i suit une loi de Bernoulli de paramètre noté p
3. $H_0 : p = 1/2$
4. règle de décision:
   - $T = \bar{X}$;
   - $\mathcal{R} = \{t : |t - 1/2| > l\}$, où l est un nombre fixé;

Rejet de H_0 si $|\bar{X} - 1/2| > l$

NB: d'autres $\mathcal{R}$ sont possibles, par exemple $\mathcal{R} = \{t : t - 1/2 > l\}$

Exemple: production de pommes

1. données: production $x_1, x_2, \ldots x_n$ d'un échantillon d'arbres de la nouvelle variété
2. modèle: $X \sim \mathcal{N}(\mu, \sigma^2)$, avec σ^2 connu
3. H_0: $\mu = 12$ (production de l'ancienne variété)
4. règle de décision:
   - $T = \bar{X}$;
   - $\mathcal{R} = \{t : t - 12 > l\}$, où l est un nombre fixé;

Rejet de H_0 si $x - 12 > l$

NB: d'autres $\mathcal{R}$ sont possibles, par exemple $\mathcal{R} = \{t : |t - 12| > l\}$

Exemple: diamètre des pommes

1. données: diamètres $x_1, x_2, \ldots x_n$ d'un échantillon de pommes de la nouvelle variété
2. modèle $X \sim \mathcal{N}(\mu, \sigma^2)$
3. H_0: $\sigma = 1$ (paramètre de l'ancienne variété)
4. règle de décision:
   - $T = S$, où $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{X})^2$
   - $\mathcal{R} = \{t : t < 1 - l\}$, où l est un nombre fixé;

Rejet de H_0 si $S < 1 - l$

NB: d'autres $\mathcal{R}$ sont possibles, par exemple $\mathcal{R} = \{t : |t - 1| > l\}$

Description des hypothèses

H_0 : hypothèse nulle

C'est l'hypothèse privilégiée. C'est celle que l'on garde si le résultat de l'expérience n'est pas clair

Analogie avec la justice:

- l'accusé est présumé innocent et c'est à l'accusation d'apporter la preuve de sa culpabilité
- "accepter" H_0, c'est acquitter faute de preuves

H_1 : hypothèse alternative

On appelle **hypothèse alternative** notée H_1 une hypothèse différente de H_0 (souvent, son contraire)

Les rôles de H_0 et H_1 ne sont donc pas du tout symétriques

Différentes formes de H_0 et H_1

Hypothèses simples

$H_0 : \{\theta = \theta_0\}$ et $H_1 : \{\theta = \theta_1\}$

Test unilatéral

- unilatéral à droite: $H_0 : \{\theta = \theta_0\}$ et $H_1 : \{\theta > \theta_0\}$
- unilatéral à gauche: $H_0 : \{\theta = \theta_0\}$ et $H_1 : \{\theta < \theta_0\}$

Test bilatéral

$H_0 : \{\theta = \theta_0\}$ et $H_1 : \{\theta \neq \theta_0\}$

Exemples

GC

- $H_0 : p = 1/2, H_1 : p > 1/2$
- $H_0 : p = 1/2, H_1 : p \neq 1/2$

Production de pommes

- $H_0 : \mu = 12, H_1 : \mu > 12$
- $H_0 : \mu = 12, H_1 : \mu \neq 12$

Diamètre des pommes

- $H_0 : \sigma = 1, H_1 : \sigma < 1$
- $H_0 : \sigma = 1, H_1 : \sigma \neq 1$

Le choix de H_1 dicte la forme de la région de rejet $\mathcal{R}$

Niveau et puissance d'un test

Quatre issues possibles pour un test

		Décision	
		H_0	H_1
Vérité	H_0	Décision correcte	Erreur de première espèce
	H_1	Erreur de seconde espèce	Décision correcte

Risque de type I

Proba de rejeter H_0 alors qu'elle est vraie

Niveau d'un test (noté α)

Plus grande valeur possible du risque I (parmi toutes les valeurs possibles sous H_0)

Pour une hypothèse nulle simple, Risque I = niveau

Risque de type II

Proba de ne pas rejeter H_0 alors qu'elle est fausse

Puissance du test (notée $1 - \beta$)

Probabilité de rejeter H_0 alors qu'elle est fausse

Puissance + Risque II = 1

Niveau et puissance: exemple

Dans un test de dépistage, les hypothèses sont

- H_0 : le patient n'est pas atteint
- H_1 : : le patient est atteint

Risque I: probabilité de déclarer atteints des patients sains (ou faux positifs)

Le niveau est la proportion de patients sains mal classés

Risque II: probabilité de déclarer sains des patients atteints

La puissance est la proportion de malades bien détectés

Remarque

Si l'échantillon reste inchangé, une diminution de α entraîne une augmentation de β et inversement. Autrement dit, si on décide de réduire le nombre de faux positifs, on augmente forcément le nombre de faux négatifs. La seule manière d'améliorer les deux critères est d'augmenter la taille de l'échantillon.

Construction de la règle de décision

La règle est construite pour que le **niveau du test soit égal à une valeur fixée *a priori***

Interprétation

On contrôle le **risque de rejeter H_0 à tort**

La règle de décision ne dépend que de *ce qui se passe sous H_0*

Exemples

- GC, $H_0 : p = 1/2, H_1 : p > 1/2$
  $\rightarrow$ on contrôle le risque de déclarer que $p > 1/2$ alors qu'en réalité $p = 1/2$
- Production de pommes, $H_0 : \mu = 12, H_1 : \mu > 12$
  $\rightarrow$ on contrôle le risque de déclarer que $\mu > 12$ alors qu'en réalité $\mu = 12$
- Diamètre des pommes, $H_0 : \mu = 12, H_1 : \sigma < 1$
  $\rightarrow$ on contrôle le risque de déclarer que $\sigma < 1$ alors qu'en réalité $\sigma = 1$

Exemple: production de pommes

On suppose que la productivité des arbres suit une loi $\mathcal{N}(\mu, 1)$
On sait alors que pour un échantillon de n arbres, la loi de $\bar{X}$ est $\mathcal{N}(\mu, 1/n)$.

Mise en place du test

- Test de H_0: "$\mu = 12kg$" contre H_1: "$\mu > 12kg$"
- Niveau choisi: $\alpha = 0.05$
- Statistique de test: $\bar{X}$ pour $n = 5$
- **Sous** H_0, $\bar{X}$ suit une loi $\mathcal{N}(12, 0.2)$, **connue**
- Forme du test: rejet de $H_0 \Leftrightarrow \bar{X} \geq u$
- Seuil de rejet: $u = 12.7356009$

Réalisation du test

- Données: 12.5, 13.4, 17, 10.7, 11.4
- Statistique de test: 13

Conclusion

Au vu de cette expérience, au niveau 0.05, on rejette l'hypothèse selon laquelle la productivité moyenne d'un arbre de la nouvelle variété est 12kg

Influence du niveau du test

Autres réalisations ($\alpha = 0.05$)

	STATISTIQUE	DECISION
1	12.5	non rejet
2	13	rejet
3	13.5	rejet

Autres réalisations ($\alpha = 0.01$)

	STATISTIQUE	DECISION
1	12.5	non rejet
2	13	non rejet
3	13.5	rejet

Illustration graphique

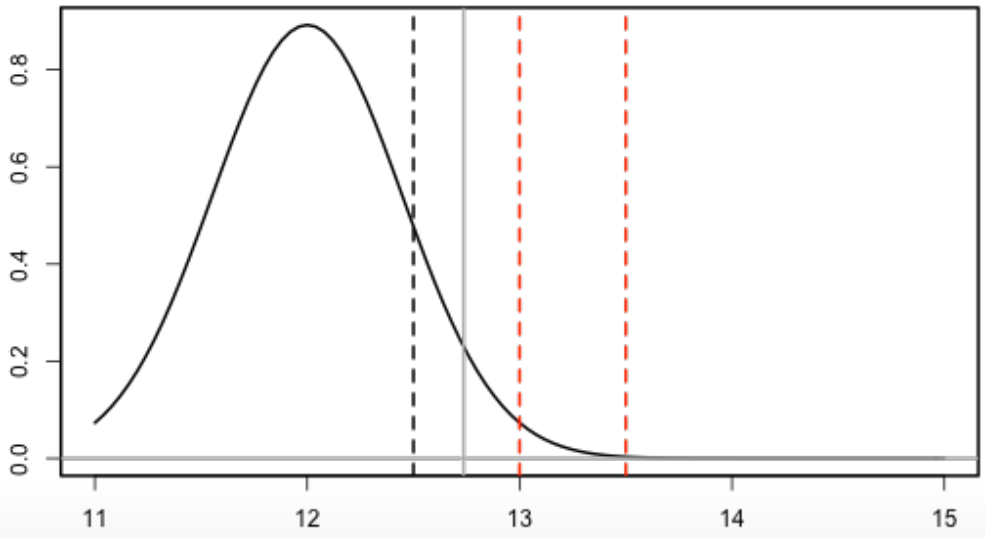
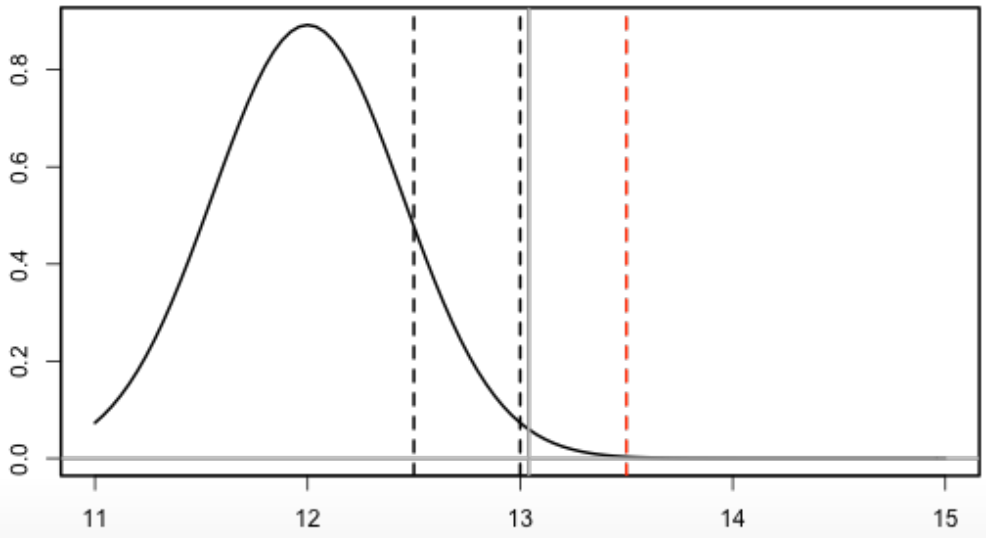


Illustration graphique



Probabilité critique

Egalement appelée p-valeur, p-value, ou degré de signification

Motivation

Enrichir la décision binaire "rejet" ou "non-rejet"

Définition

Probabilité de rejeter à tort l'hypothèse nulle

Niveau et p-valeur

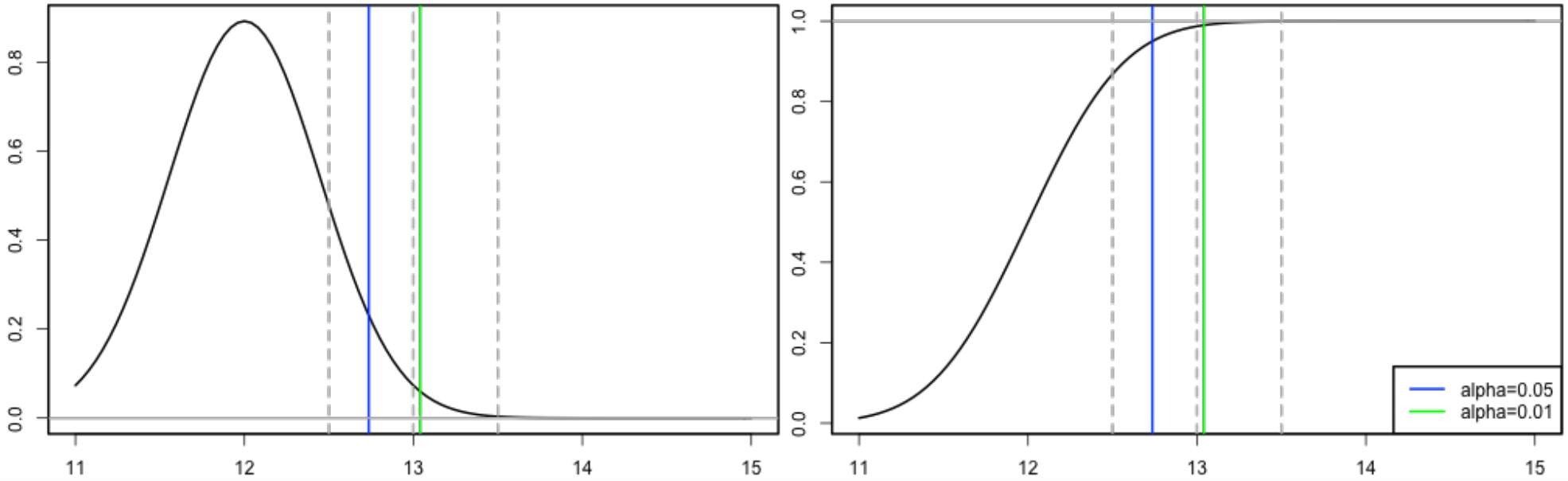
Il y a équivalence entre:

- la p-valeur est inférieure à α
- l'hypothèse nulle est rejetée au niveau α

Exemple: production de pommes

	STATS	P-VALUE	DECISION (0.05)	DECISION (0.01)
1	12.5	0.13	non rejet	non rejet
2	13	0.013	rejet	non rejet
3	13.5	4e-04	rejet	rejet

Illustration graphique



Tests de comparaison

à une valeur de référence

Comparaison à une valeur de référence

1. Tests sur la moyenne

- loi normale, variance connue
- loi normale, variance inconnue
- loi quelconque, grands échantillons

2. Tests sur une proportion

Remarque

Dans tous les cas ci-dessus on peut faire un test **bilatéral** ou **unilatéral** selon la forme de H_1

Tests sur la moyenne

Loi normale et loi de Student: propriétés

Soit $X_1, X_2, \dots X_n$ un échantillon de n variables

- indépendantes
- de même loi $\mathcal{N}(\mu, \sigma^2)$

Alors $\bar{X}$ suit une loi $\mathcal{N}(\mu, \sigma^2/n)$

Ceci peut s'écrire:

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

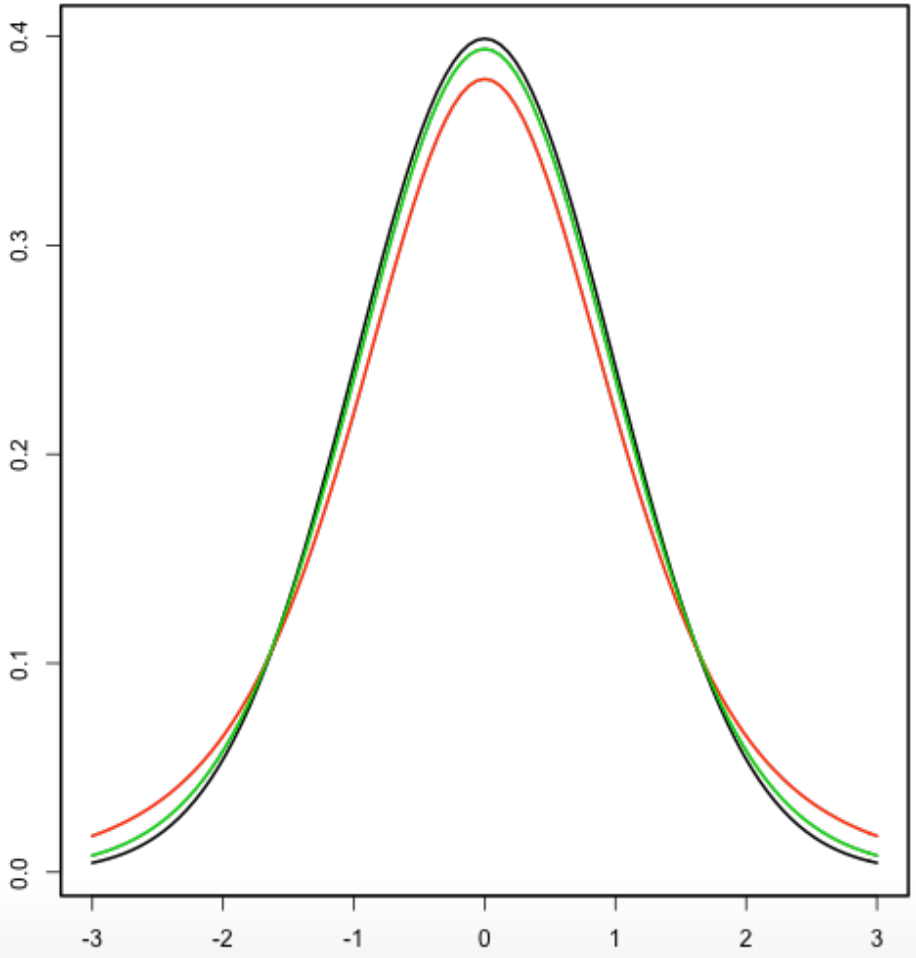
Lorsque σ est inconnu, on peut l'estimer par S. On a alors:

$$\frac{\bar{X} - \mu}{S} \sim t_{n-1}$$

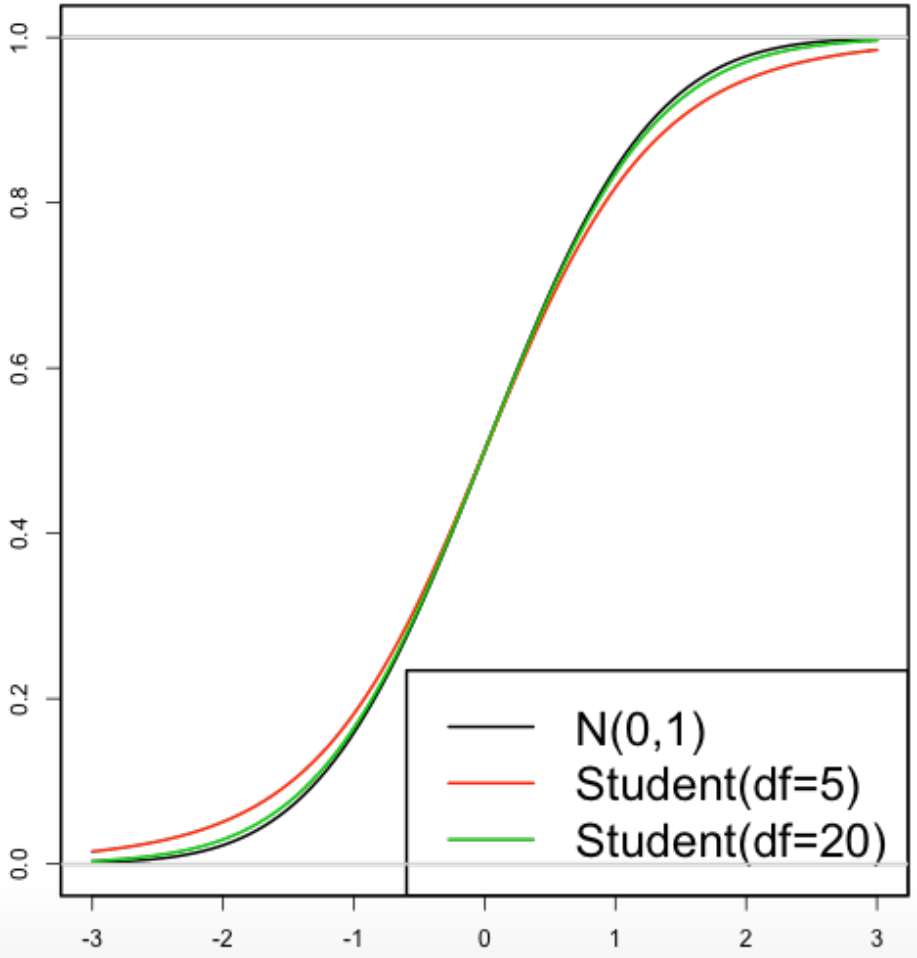
où t_{n-1} désigne la **loi de Student** à $n - 1$ degrés de liberté

Loi normale et loi de Student: illustration

Densité



Fonction de répartition



Test sur la moyenne: loi normale, variance connue

- $(X_1, X_2, \dots X_n)$ sont indépendantes, identiquement distribuées de loi $\mathcal{N}(\mu, \sigma^2)$
- μ est inconnu, σ **est connu**, μ_0 est une valeur de référence, connue
- Statistique de test: $T = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$
- Sous $H_0 : \mu = \mu_0$, T suit une loi $\mathcal{N}(0, 1)$

Test bilatéral de $H_0 : \mu = \mu_0$ contre $H_1 : \mu \neq \mu_0$

- Test de niveau α: rejeter H_0 si $|t| \geq u$,

où u est le quantile d'ordre $1 - \alpha/2$ de $\mathcal{N}(0, 1)$

Exemple: si $\alpha = 0.05$ alors $u = 1.959964$

- p-value du test: $p = \mathrm{P}_{H_0}(|T| \geq |t|)$

Test sur la moyenne: loi normale, variance inconnue

- $(X_1, X_2, \dots X_n)$ sont indépendantes, identiquement distribuées de loi $\mathcal{N}(\mu, \sigma^2)$
- μ est inconnu, σ **est inconnu**, μ_0 est une valeur de référence, connue
- Statistique de test: $T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$
- Sous $H_0 : \mu = \mu_0$, T suit une loi de Student à $n - 1$ degrés de liberté

Test bilatéral de $H_0 : \mu = \mu_0$ contre $H_1 : \mu \neq \mu_0$

- Test de niveau α: rejeter H_0 si $|t| \geq u$,

où u est le quantile d'ordre $1 - \alpha/2$ de la loi de Student à $n - 1$ degrés de liberté

Exemples: si $\alpha = 0.05$, alors

- (i) pour $n = 20$, $u = 2.0930241$;
- (ii) pour $n = 200$, $u = 1.9719565$

  - p-value du test: $p = \mathrm{P}_{H_0}(|T| \geq |t|)$

Test sur la moyenne: loi normale, variance connue

- $(X_1, X_2, \dots X_n)$ sont indépendantes, identiquement distribuées de loi $\mathcal{N}(\mu, \sigma^2)$
- μ est inconnu, σ **est connu**, μ_0 est une valeur de référence, connue
- Statistique de test: $T = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$
- Sous $H_0 : \mu = \mu_0$, T suit une loi $\mathcal{N}(0, 1)$

Test unilatéral (à gauche)

- $H_0 : \mu = \mu_0$ contre $H_1 : \mu < \mu_0$
- Test de niveau α: rejet de H_0 si $t \leq u$,

où u est le quantile d'ordre α de $\mathcal{N}(0, 1)$

Si $\alpha = 0.05$, alors $u = -1.645$

- p-value du test: $p = \mathrm{P}_{H_0}(T \leq t)$

Test unilatéral (à droite)

- $H_0 : \mu = \mu_0$ contre $H_1 : \mu > \mu_0$
- Test de niveau α: rejet de H_0 si $t \geq u$,

où u est le quantile d'ordre $1 - \alpha$ de $\mathcal{N}(0, 1)$

Si $\alpha = 0.05$, alors $u = 1.645$

- p-value du test: $p = \mathrm{P}_{H_0}(T \geq t)$

Test sur la moyenne: loi normale, variance inconnue

- $(X_1, X_2, \dots X_n)$ sont indépendantes, identiquement distribuées de loi $\mathcal{N}(\mu, \sigma^2)$
- μ est inconnu, σ **est inconnu**, μ_0 est une valeur de référence connue
- Statistique de test: $T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$
- Sous $H_0 : \mu = \mu_0$, T suit une loi de Student à $n - 1$ degrés de liberté

Test unilatéral (à gauche)

- $H_0 : \mu = \mu_0$ contre $H_1 : \mu < \mu_0$
- Test de niveau α: rejet de H_0 si $t \leq u$,

où u est le quantile d'ordre α de t_{n-1}

Si $\alpha = 0.05$, alors $u = -1.729$

- p-value du test: $p = \mathrm{P}_{H_0}(T \leq t)$

Test unilatéral (à droite)

- $H_0 : \mu = \mu_0$ contre $H_1 : \mu > \mu_0$
- Test de niveau α: rejet de H_0 si $t \geq u$,

où u est le quantile d'ordre $1 - \alpha$ de t_{n-1}

Si $\alpha = 0.05$, alors $u = 1.729$

- p-value du test: $p = \mathrm{P}_{H_0}(T \geq t)$

Applicabilité

Les tests précédents sont valides lorsque les variables $(X_1, X_2, \dots X_n)$ sont **indépendantes**, identiquement distribuées **de loi** $\mathcal{N}(\mu, \sigma^2)$

Robustesse à la non-normalité

Sauf raison objective d'écarter l'hypothèse de normalité (de X), on considère généralement que ces tests sont applicables dès que n est "assez grand" (disons $n \geq 30$).

Il existe des tests ne nécessitant pas de faire d'hypothèse de normalité:

- tests reposant sur le TLC, **uniquement valables pour "n grand"** (pages suivantes)
- **tests non-paramétriques** (mentionnés dans les chapitres suivants)

Non-robustesse aux écarts à l'indépendance

En revanche ces tests ne sont **pas robustes** à un écart à l'hypothèse d'indépendance

Test sur la moyenne: loi quelconque, n grand

- $(X_1, X_2, \dots X_n)$ sont indépendantes, identiquement distribuées de loi $\mathcal{L}$ quelconque
- $\mu = E(X_1)$ est inconnu, μ_0 est une valeur de référence, connue
- Statistique de test: $T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$
- Sous $H_0 : \mu = \mu_0$, T suit **approximativement** une loi $\mathcal{N}(0, 1)$

Test bilatéral de $H_0 : \mu = \mu_0$ contre $H_1 : \mu \neq \mu_0$

- Test de niveau α: rejeter H_0 si $|t| \geq u$,

où u est le quantile d'ordre $1 - \alpha/2$ de la loi $\mathcal{N}(0, 1)$

Exemples: si $\alpha = 0.05$, alors $u = 1.96$;

- p-value du test: $p = \mathrm{P}_{H_0}(|T| \geq |t|)$

Test sur la moyenne: loi quelconque, n grand

- $(X_1, X_2, \dots X_n)$ sont indépendantes, identiquement distribuées de loi $\mathcal{L}$ quelconque
- $\mu = E(X_1)$ est inconnu, μ_0 est une valeur de référence, connue
- Statistique de test: $T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$
- Sous $H_0 : \mu = \mu_0$, T suit **approximativement** une loi $\mathcal{N}(0, 1)$

Test unilatéral (à gauche)

- $H_0 : \mu = \mu_0$ contre $H_1 : \mu < \mu_0$
- Test de niveau α: rejet de H_0 si $t \leq u$,

où u est le quantile d'ordre α de $\mathcal{N}(0, 1)$

Si $\alpha = 0.05$, alors $u = -1.645$

- p-value du test: $p = \mathrm{P}_{H_0}(T \leq t)$

Test unilatéral (à droite)

- $H_0 : \mu = \mu_0$ contre $H_1 : \mu > \mu_0$
- Test de niveau α: rejet de H_0 si $t \geq u$,

où u est le quantile d'ordre $1 - \alpha$ de $\mathcal{N}(0, 1)$

Si $\alpha = 0.05$, alors $u = 1.645$

- p-value du test: $p = \mathrm{P}_{H_0}(T \geq t)$

Tests sur les proportions

Loi de Bernoulli et loi binomiale: définitions

Soit $X_1, X_2, \dots X_n$ un échantillon de n variables

- indépendantes
- de même loi $B(p)$: de Bernoulli de paramètre p

Alors $\sum_{i=1}^{n} X_i$ suit une loi binomiale $Bin(n, p)$

Remarque: si n est grand alors la loi de $\bar{X}$ se rapproche d'une loi normale de paramètres $\mathcal{N}(p, \frac{p(1-p)}{n})$

Test sur une proportion

- $(X_1, X_2, \dots X_n)$ sont indépendantes, identiquement distribuées de loi $B(p)$
- p est inconnu, p_0 est une valeur de référence, connue
- Statistique de test: $T = \sum_{i=1}^{n} X_i$
- Sous $H_0 : p = p_0$, T suit une loi $Bin(n, p_0)$

Test bilatéral de $H_0 : p = p_0$ contre $H_1 : p \neq p_0$

- Test de niveau α: rejeter H_0 si $t \leq u_g$ ou $t \geq u_d$,

où u est le quantile d'ordre $1 - \alpha/2$ de la loi $Bin(n, p_0)$

Exemple: si $\alpha = 0.05$, $n = 10$ et $p_0 = 0.5$, alors $u_g = 2$ et $u_d = 8$

- p-value du test: $p = 2 \min(p_g, 1 - p_g)$, où $p_g = \mathrm{P}_{H_0}(T \leq t)$

(la loi $Bin(n, p_0)$ n'est pas symétrique par rapport à 0 donc la formule pour la p-valeur est inhabituelle)

Test sur une proportion

- $(X_1, X_2, \dots X_n)$ sont indépendantes, identiquement distribuées de loi $B(p)$
- p est inconnu, p_0 est une valeur de référence, connue
- Statistique de test: $T = \sum_{i=1}^{n} X_i$
- Sous $H_0 : p = p_0$, T suit une loi $Bin(n, p_0)$

Test unilatéral (à gauche)

- $H_0 : p = p_0$ contre $H_1 : p < p_0$
- Test de niveau α: rejet de H_0 si $t \leq u$,

où u est le quantile d'ordre α de $Bin(n, p_0)$

Si $\alpha = 0.05$, alors $u = 2$

- p-value du test: $p = \mathrm{P}_{H_0}(T \leq t)$

Test unilatéral (à droite)

- $H_0 : p = p_0$ contre $H_1 : p > p_0$
- Test de niveau α: rejet de H_0 si $t \geq u$,

où u est le quantile d'ordre $1 - \alpha$ de $Bin(n, p_0)$

Si $\alpha = 0.05$, alors $u = 8$

- p-value du test: $p = \mathrm{P}_{H_0}(T \geq t)$