Premières notions de satistique
Introduction aux tests statistiques

Franck Picard

UMR CNRS-5558, Laboratoire de Biométrie et Biologie Evolutive

franck.picard@univ-lyon1.fr

La statistique une science citoyenne?

- 49/51, la gauche gagne ! quelle confiance accordez vous à cette affirmation ?
- Les OGM sont dangereux pour la santé ! c'est sûr ?
- La population que j'observe est-elle à l'équilibre d'Hardy Weinberg ?
- Y-a-t-il une proportion plus élevée de suicide dans mon entreprise que dans la population générale ?
- La terre se réchauffe ?
- Fumer tue

COUPABLE OU NON-COUPABLE ?

Tests et démarche scientifique

- Après avoir estimé un paramètre, que conclure de sa valeur ?
- Par exemple: la proportion estimée d'électeurs qui ont voté A est $\widehat{p}(\mathbf{x}) = 0.45$. Que peut-on en dire ?
- L'estimation consiste on cherche à collecter des informations sur un paramètre, et à l'estimer au mieux
- La démarche des tests consiste à **comparer** le résultat observé avec une valeur de référence: la proportion d'électeur est-elle supérieure ou égale à 50%?
- Répondre à cette question suppose que l'on prenne en compte la variabilité de l'échantillon.

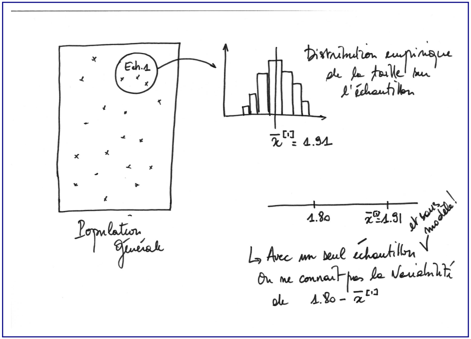
La motivation des tests est la prise de décision à partir de résultats aléatoires

Une ? Deux ? Plusieurs populations ?

- Si on ne dispose que d'une seule population, on compare en général le résultat à une référence:
  - Pour un *paramètre*: 0 (moyenne), 0.5 (proportion)
  - Pour une *distribution*: comparer aux distribution connues: les données sont-elles distribuées comme une loi gaussienne ?
- Lorsque l'on observe deux populations, on cherche souvent à les comparer
  - La moyenne de la taille des filles et des garcons est-elle la même ?
  - La répartition du QI est-elle la même pour les filles et les garcons ?
- Les modèles linéaires (ANOVA) permettront d'étendre ces démarches au cas de plusieurs populations

Tester sans modèle ?

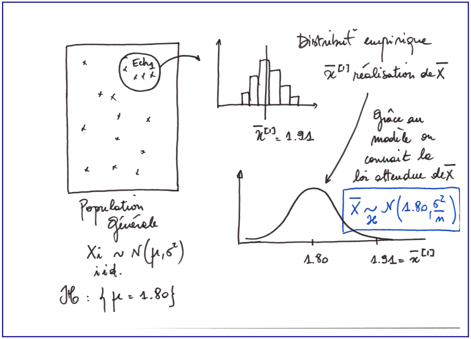
- On enregistre la taille de n individus notées $x_1, ...x_n$. On estime la taille moyenne de l'**échantillon** avec $\overline{x} = 1.91$m
- On souhaite déterminer si la taille des individus de la **population** est égale à 1.80m en moyenne.
- Sans modèle on peut calculer $\overline{x} - 1.80$ et conclure que la taille de la population d'intérêt est différente de 1.80m.



Mais cette comparaison ne prend pas en compte la variabilité des données (fluctuations d'échantillonnage, exemple des arbres).

Le cadre des tests paramétriques

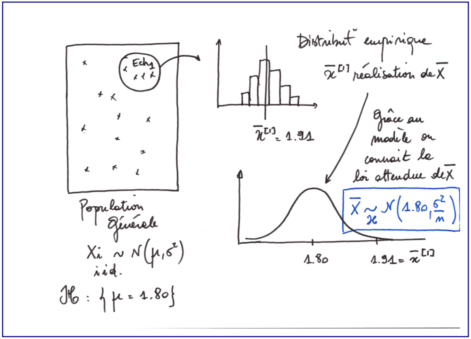
- La distribution des données est **modélisée** en utilisant une loi qui dépend d'un paramètre (notée F_θ)
- Exemple de la taille (on suppose σ^2 connue) $X_i \sim \mathcal{N}(\mu, \sigma^2)$ (iid).
- Test **paramétrique**: qui concerne les paramètres du modèle.
- L'hypothèse posée concerne toujours les **paramètres** du modèle (μ et pas $\overline{X}$)



Ici on pose l'hypothèse $H_0 : \{\mu = 1.80\}$

Modèle et loi de la statistique-1

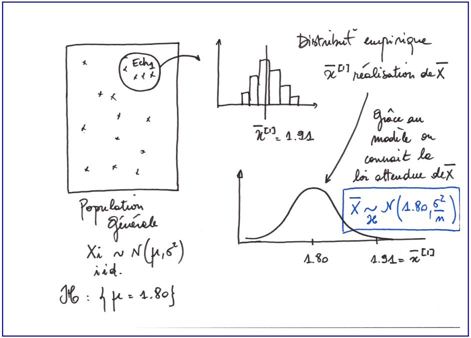
- On appelle statistique de test une fonction des observations $\mathbf{x}$ qui permet de tester l'hypothèse H_0
- On la construit pour que sa distribution sous H_0 soit connue
- La statistique de test est en lien avec l'estimateur du paramètre du modèle
- Ici: $T(\mathbf{x}) = \overline{x}$ ou $T(\mathbf{x}) = \overline{x}/\sigma$ (σ connue)



Si on ne dispose pas d'un modèle sur les observations X_i alors on ne peut pas quantifier la variabilité de la statistique de test $T(\mathbf{x})$.

Modèle et loi de la statistique-2

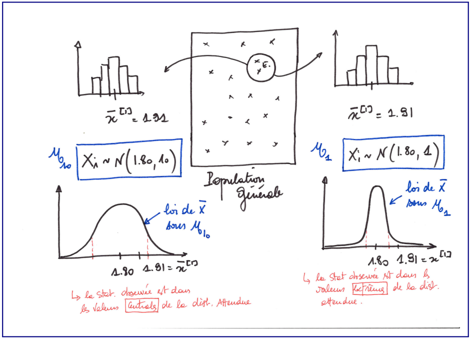
- On pose un modèle sur les X_i (par exemple X_i, iid, $\mathcal{N}(\mu, \sigma^2)$)
- La statistique de test $T(\mathbf{x})$ est **une** réalisation de $T(\mathbf{X})$
- Etant donné que l'on a **choisi** une loi pour les X_i on peut en déduire une loi pour $T(\mathbf{X})$



Le modèle permet de quantifier la variabilité de la statistique de test, par exemple $T(\mathbf{X}) = \overline{X} \underset{H_0}{\sim} \mathcal{N}(1.80, \sigma^2/n)$

Modèle et loi de la statistique-3

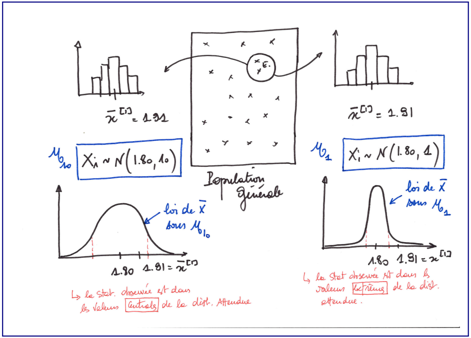
- Sous l'hypothèse du modèle $\mathcal{N}(1.80, \sigma^2)$, les quantiles de la loi normales donnent des intervalles **prévus** de variations de $\overline{X}$.
- Si on prévoit un modèle avec plus de dispersion, il faudra un écart de moyenne plus important pour détecter une différence atypique



La conclusion d'un test paramétrique dépend essentiellement du modèle posé sur les observations

Vers le retour des quantiles

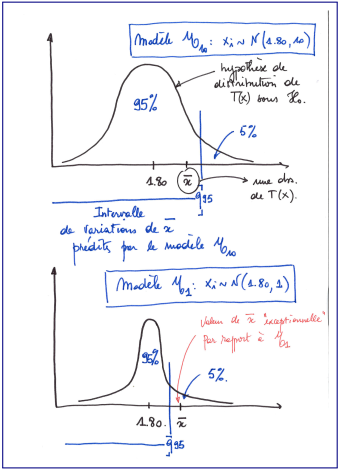
- Si $\overline{x}$ est dans les valeurs "médianes" de la gaussienne alors on pourra dire que la probabilité que la taille de la population soit de 1.80 est forte
- Si $\overline{x}$ est dans les valeurs "extrêmes" de la gaussienne alors on pourra dire que la probabilité que la taille de la population soit égale à 1.80 est faible



Les quantiles sont utilisés pour positionner la valeur observée $T(\mathbf{x})$ de la statistique par rapport à la distribution attendue de $T(\mathbf{X})$ sous H_0

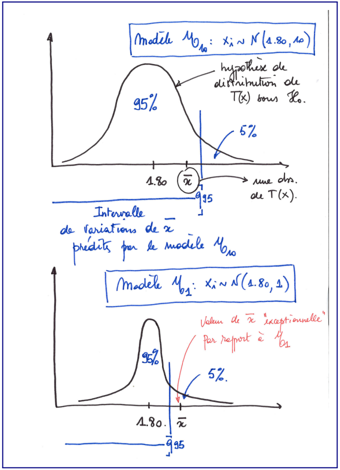
Quantiles et Zone de rejet

- Si $\overline{x} \in]-\infty, q_{1-\alpha}]$ alors on acceptera H_0
- Si $\overline{x} \in [q_{1-\alpha}, \infty[$ alors on rejettera H_0
- La zone de rejet définit l'ensemble des valeurs de $T(\mathbf{x})$ pour lesquelles on rejette H_0
- On notera α la part attendue sous H_0 des valeurs de $T(\mathbf{X})$ dans la zone de rejet



Règle de décision

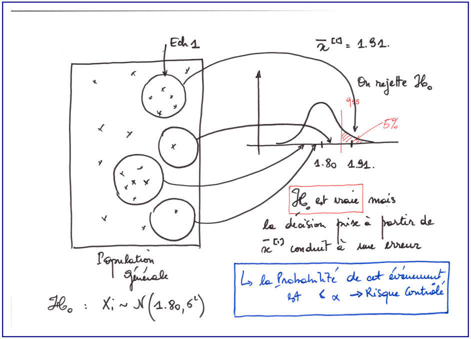
- La démarche fondamentale consiste à supposer que l'hypothèse nulle H_0 est vérifiée
- Le raisonnement consiste à s'interroger sur le caractère plausible ou non de l'observation de $T(\mathbf{x})$ sous cette hypothèse
- La procédure consiste à **rejeter** H_0 quand $T(\mathbf{X})$ **dépasse** un seuil



$$\{\text{Rejet de } H_0\} \Longleftrightarrow \{T(\mathbf{X}) \geq \text{seuil}\}$$

Pourquoi choisir les quantiles comme seuil ?

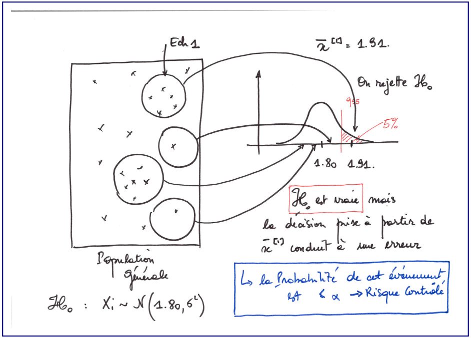
- Les quantiles permettent de quantifier $\mathbb{P}_0\{T(\mathbf{X}) \geq q_{1-\alpha}\} \leq \alpha$
- C'est la probabilité sous H_0 qu'un échantillon donne une taille qui dépasse $q_{1-\alpha}$
- Si on tirait un autre échantillon et que l'on refaisait une mesure de $T(\mathbf{X})$, on n'aurait que $\alpha\%$ de "chance" que cette nouvelle mesure dépasse $q_{1-\alpha}$



$\mathbb{P}_0\{T(\mathbf{X}) \geq q_{1-\alpha}\}$ est la masse "résiduelle" de distribution de $T(\mathbf{X})$ qu'il resterait si on rejetait H_0 à partir de $q_{1-\alpha}$

Notion de risque de première espèce

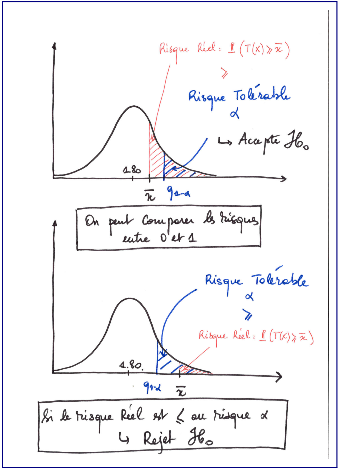
- Le principe des tests est de prendre une décision
- Donc le principe des tests est de faire des erreurs
- L'avantage des statistiques est de pouvoir quantifier ces erreurs
- Si on choisit $q_{1-\alpha}$ comme seuil, alors on a une probabilité de α de rejeter alors que l'hypothèse est vraie



Le risque de première espèce correspond à la probabilité d'avoir un faux positif

Degré de significativité

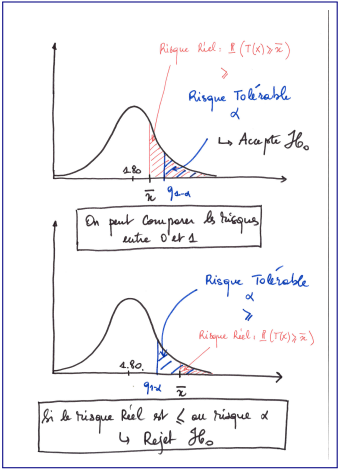
- Besoin de calculer les quantiles de la loi de $T(\mathbf{X})$ sous H_0
- Besoin de recalculer le quantile si on change α
- On cherche alors à **quantifier le degré de significativité de la décision**
- $\mathbb{P}_0\{T(\mathbf{X}) \geq t\}$ quantifie la "queue" de distribution de la statistique de test.



$\forall t,\ \mathbb{P}_0\{T(\mathbf{X}) \geq t\}$: c'est le risque pris en rejettant H_0 à partir de t

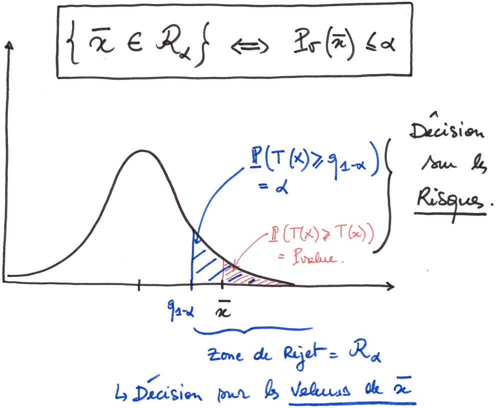
Définition de la P-valeur

- C'est la quantité qui est utilisée par tous les logiciels pour prendre une décision
- La quantité $\mathbb{P}_0\{T(\mathbf{X}) \geq T(\mathbf{x})\}$ quantifie le risque que l'on prend en rejetant l'hypothèse avec les données observées $\mathbf{x}$
- On la note $P_v(\mathbf{x})$, c'est un risque réel que l'on compare à un risque admissible
- Pour contrôler le risque α **la règle de Décision** sera:



$$\{\mathbb{P}_0\{T(\mathbf{X}) \geq T(\mathbf{x})\} \leq \alpha\} \text{ on rejette } H_0.$$

Deux règles de décision équivalentes



Résumé de la procédure de test

1. On recueille des données $(x_1, ...x_n)$
2. On modélise les observations $(X_1, \dots, X_n)$ à l'aide d'un modèle de distribution F_θ
3. On définit une hypothèse nulle à tester H_0
4. On définit une statistique de test $T(\mathbf{X})$ pour tester H_0 et on l'évalue sur l'échantillon $T(\mathbf{x})$
5. On calcule la probabilité de dépassement sous H_0 $\mathbb{P}_0\{T(\mathbf{X}) \geq T(\mathbf{x})\}$ c'est la p-valeur ou "p-value"
6. On fixe un risque α
7. Si $\{\mathbb{P}_0\{T(\mathbf{X}) \geq T(\mathbf{x})\} \leq \alpha\}$ on rejette H_0

Outline

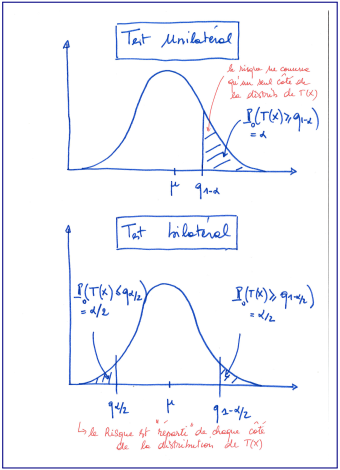
Définition de plusieurs hypothèses. L'hypothèse nulle

- On commence par définir l'hypothèse nulle H_0: c'est l'hypothèse que l'on souhaite tester
- Exemple: on observe 99 votes pour `"x"` sur 100, est ce que `"x"` gagne ? On posera $H_0 : \{p = 0.5\}$
- Un principe des tests est que l'hypothèse nulle correspond à l'absence d'effet (ou de signal).
- La démarche consiste à accumuler des données pour rejeter cette hypothèse. H_0 est l'hypothèse à réfuter.

Sous H_0 on est présumé innocent.

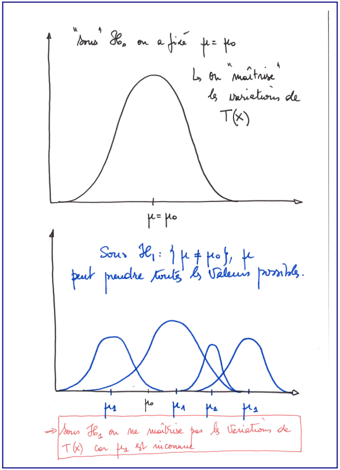
Définition de plusieurs hypothèses. L'hypothèse alternative

- C'est l'hypothèse "contre" laquelle on teste l'hypothèse nulle.
- Elle est en général définie par un/des intervalles
- Si H_1 ne concerne qu'une partie de la distribution de $T(\mathbf{X})$ alors le test est un test uni-latéral. Ex: $H_1 : \{\mu > 1.80\}$
- Si H_1 concerne les deux parties "extrêmes" de la distribution de $T(\mathbf{X})$ alors le test est un test bi-latéral. Ex: $H_1 : \{\mu \neq 1.80\}$



Pourquoi se placer “sous” H_0

- H_0 est une hypothèse de travail
- En supposant qu'elle est vérifiée, on sait dériver les caractéristiques de $T(\mathbf{X})$
- Sous H_1 au contraire, on ne sait rien. On sait simplement que le paramètre est différent de la valeur supposée sous H_0



Les deux types d'erreur

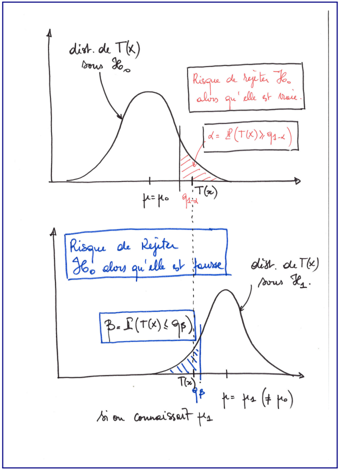
- L'erreur de première espèce α

  $\mathbb{P}\{\text{Décider } H_1 \text{ alors que } H_0 \text{ est vraie}\}$

- L'erreur de deuxième espèce β

  $\mathbb{P}\{\text{Décider } H_0 \text{ alors que } H_1 \text{ est vraie}\}$

- La puissance d'un test:
  $\pi = 1 - \beta$
- La determination de π dépend de ce qui se passe sous H_1 (souvent inaccessible)



On se trompe toujours ! Choix de l'absurde

- Les deux risques sont liés et varient généralement en sens inverse:

MAIS l'idée de Neyman et Pearson est de supposer que **les hypothèses H_0 et H_1 ne jouent pas des rôles symétriques**

- En général H_0 suppose l'**absence d'effet**

La stratégie consiste à fixer un risque tolérable α (faux positifs), et de trouver le test qui maximise la puissance π

Du rôle central de la définition de l'hypothèse nulle

- Fixer α *a priori* correspond au **principe de précaution**
- Plus α diminue, plus le test devient **conservatif**: on aura tendance à conserver H_0

Cet a priori signifie que α est le risque maximum que l'on est prêt à prendre en rejetant H_0 à tort

Outline

Les principaux tests à connaître

- Comparaison d'**un paramètre** (espérance, probabilité de succès, variance) à une valeur de référence (test gaussien, de Student, binomial, et du chi2, test de rang)
- Comparaison d'**une distribution** empirique à une distribution théorique de référence (test du chi2, de Kolmogorov-Smirnov)
- Comparaison de **deux populations** (espérances, probabilités de succès, variance), tests gaussiens, de Student, binomial et de Fisher
- Comparaison de **deux distributions** (Kolmogorov, test de rang)
- Test d'indépendance

...Ou comment s'y retrouver ?

- La diversité des situations, et l'inventivité des statisticiens crééent une diversité de situations / tests possibles
- En pratique, la difficulté est souvent: "Je fais quoi dans quelle situation" ?

En statistique, on raisonne (toujours) en terme d'information disponible: rôle central du nombre d'observations

- On aura souvent la contrainte du nombre d'observations disponibles (réalité expérimentale)
- Le fil directeur du choix utilise un principe simple:

Plus on dispose d'information, plus on peut faire des hypothèses fortes

Hypothèses fortes / faibles pour comparer deux populations

- La contrainte provient principalement de la disponibilité des données
- Le caractère fort/faible des hypothèses concerne essentiellement la spécification du modèle
- Si F_θ spécifie une loi particulière: on fait une hypothèse très forte sur la distribution des données et sur sa paramétrisation.

Dans ce cas, il faut "beaucoup" d'observations, et on se focalise **sur le paramètre** $\{\theta_0 = \theta_1\}$

- Si on a moins d'information, peut-être que la distribution des observations ne peut être "contrainte" par un F_θ particulier

Dans ce cas, on se focalise "uniquement" **sur les distributions** $\{F_0 = F_1\}$

Outline

Est ce qu'on abbat les arbres ou pas ?

- L'exploitant de parcelles d'arbres doit décider s'il abbat ou non les arbres d'une parcelle.
- Au vu d'expertises antérieures, il sait qu'il peut abbattre les arbres quand leur taille est au moins de 25cm. Il recueille donc la taille de 11 arbres de la parcelle 1 et 10 arbres de la parcelle 2.

Type 1	23.4	24.4	24.6	24.9	25	26.2	26.3	26.8	26.8	26.9	27
Type 2	22.5	22.9	23.74	24.0	24.4	24.5	25.3	26	26.4	26.7	.

- On considère des échantillons de **2 populations indépendantes** de taille n_1 et n_2, notés $\mathbf{x}^1 = (x_1^1, \ldots, x_{n_1}^1)$ et $\mathbf{x}^2 = (x_1^2, \ldots, x_{n_2}^2)$.
- On suppose que la variable d'intérêt peut être modélisée par une **loi gaussienne**, telle que:

$$X_i^1 \underset{iid}{\sim} \mathcal{N}(\mu_1, \sigma^2), \; X_i^2 \underset{iid}{\sim} \mathcal{N}(\mu_2, \sigma^2), \; \sigma^2 \text{connue.}$$

Est ce qu'on abbat les arbres ou pas ?

H_0 On souhaite tester les hypothèses **sur le paramètre d'espérance**

- On fait l'hypothèse de "pas d'effet" (donc "on n'abbat pas les arbres"):

$$H_0 : \{\mu_1 < 25\}, \qquad H_0 : \{\mu_2 < 25\},$$

H_1 On définit l'alternative:

$$H_1 : \{\mu_1 \geq 25\}, \qquad H_0 : \{\mu_2 \geq 25\}$$

- Etant donné que l'hypothèse porte sur le paramètre d'espérance, on estime les paramètres du modèle et on calcule la statistique de test

$$T(\mathbf{X}) = \frac{\overline{X} - 25}{\sigma}\sqrt{n}, \text{ (σ supposé connu dans un premier temps)}$$

Type	Moyenne	Ecart-type	nb obs	$T(\mathbf{x})$	$\mathbb{P}\{T(\mathbf{X}) \geq t(\mathbf{x})\}$
1	25.66	1.24	11	1.7653	0.053
2	24.64	1.43	10	-0.7960979	0.222

Cas unilatéral et bilatéral

- Dans le cas d'une hypothèse uni-latérale: on calcule le degré de significativité du test avec

$$H_1 : \{\mu_1 > \mu_0\}, \mathsf{P}_v(T(\mathbf{x})) = \mathbb{P}_0\{T(\mathbf{X}) > T(\mathbf{x})\}$$

$$H_1 : \{\mu_1 < \mu_0\}, \mathsf{P}_v(T(\mathbf{x})) = \mathbb{P}_0\{T(\mathbf{X}) < T(\mathbf{x})\}$$

- Dans le cas d'une hypothèse bi-latérale:

$$H_1 : \{\mu_1 \neq \mu_0\}, \mathsf{P}_v(T(\mathbf{x})) = \mathbb{P}_0\{T(\mathbf{X}) > |T(\mathbf{x})|\}$$

- Les Pvalues se calculent à l'aide des fonctions de répartition des statistiques de test (en utilisant les logiciels)

Outline

Le test gaussien de comparaison de moyennes

Obs On considère des échantillons de **2 populations indépendantes** de taille n_1 et n_2, notés $\mathbf{x}^1 = (x_1^1, \dots, x_{n_1}^1)$ et $\mathbf{x}^2 = (x_1^2, \dots, x_{n_2}^2)$.

F_θ On suppose que la variable d'intérêt peut être modélisée par une **loi gaussienne**, telle que:

$$X_i^1 \underset{iid}{\sim} \mathcal{N}(\mu_1, \sigma^2),\ X_i^2 \underset{iid}{\sim} \mathcal{N}(\mu_2, \sigma^2),\ \sigma^2\text{connue.}$$

H_0 On souhaite tester l'hypothèse **sur le paramètre d'espérance**

$$H_0 : \{\mu_1 = \mu_2\}$$

H_1 On définit l'alternative:

$$H_1 : \{\mu_1 \neq \mu_2\} \text{ ou } H_1 : \{\mu_1 > \mu_2\},\ \text{ou } H_1 : \{\mu_1 < \mu_2\}?$$

Modèle gaussien et comparaison de moyennes

$\widehat{\mu}(\mathbf{X})$ On estime les paramètres μ_1 et μ_2:

$$\widehat{\mu}_1(\mathbf{X}) = \overline{X}^1, \text{ et } \widehat{\mu}_2(\mathbf{X}) = \overline{X}^2$$

$T(\mathbf{X})$ Une statistique naturelle pour tester H_0:

$$\overline{X}^1 - \overline{X}^2 \underset{H_0}{\sim} \mathcal{N}(0, \sigma^2(1/n_1 + 1/n_2))$$

- On utilise en général la statistique centrée réduite (quand σ^2 est connue)

$$T(\mathbf{X}) = \frac{\overline{X}^1 - \overline{X}^2}{\sigma\sqrt{1/n_1 + 1/n_2}} \underset{H_0}{\sim} \mathcal{N}(0, 1)$$

Modèle gaussien et comparaison de moyennes

$T(\mathbf{x})$ On calcule $T(\mathbf{x})$ à partir des données

Pv On calcule le degré de significativité du test (ici on suppose une alternative bilatérale)

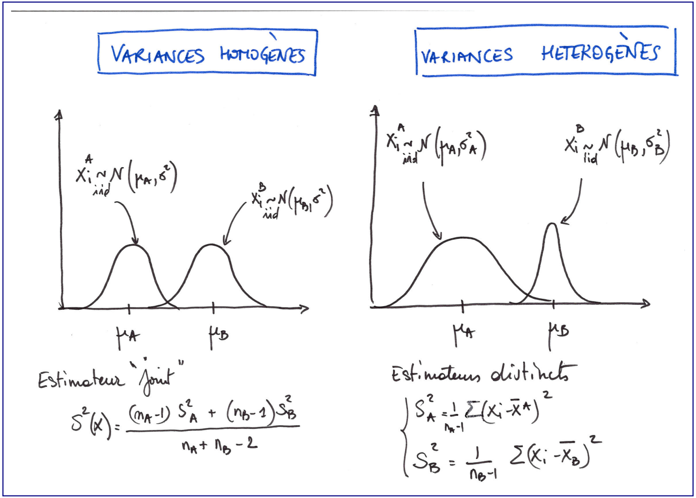
$$\mathrm{P}_v(T(\mathbf{x})) = \mathbb{P}_0\{T(\mathbf{X}) \geq |T(\mathbf{x})|\} = 1 - 2\Phi(T(\mathbf{x}))$$

- On prend une décision à l'aide de $\mathrm{P}_v(T(\mathbf{x}))$

Si $\mathrm{P}_v(T(\mathbf{x})) \leq \alpha$ on rejette H_0

★ Avant de conclure, on vérifie que le modèle n'était pas trop faux !

L'importance de bien modéliser les variances



Le test de Student à variances homogènes

- Si σ est inconnue, il faut l'estimer ce qui **change la statistique**:

$$T(\mathbf{X}) = \frac{\overline{X}^1 - \overline{X}^2}{S(\mathbf{X})\sqrt{1/n_1 + 1/n_2}}$$

- Cette statistique est constituée du ratio de deux estimateurs
- $\overline{X}^1 - \overline{X}^2$ l'estimateur de la différence de moyennes est gaussien
- L'estimateur de la variance est construit tel que:

$$S^2(\mathbf{X}) = \frac{(n_1 - 1)S^2(\mathbf{X}^1) + (n_2 - 1)S^2(\mathbf{X}^2)}{(n_1 + n_2 - 2)}$$

- C'est l'estimateur de la variance des observations regroupées ("pooled" variance)

Le test de Student à variances homogènes

- la loi du numérateur est une loi gaussienne (différence de deux moyennes de populations indépendantes)

$$\widehat{\mu}_1(\mathbf{X}) - \widehat{\mu}_2(\mathbf{X}) \underset{H_0}{\sim} \mathcal{N}(0, \sigma^2(1/n_1 + 1/n_2))$$

- la loi du dénominateur est une loi du chi2 à $n_1 + n_2 - 2$ degrés de liberté

$$S^2(\mathbf{X}) \sim \sigma^2\chi^2(n_1 + n_2 - 2)$$

- La loi d'un tel ratio est une loi de Student (admis) à $(n_1 + n_2 - 2)$ degrés de liberté

$$T(\mathbf{X}) = \frac{\overline{X}^1 - \overline{X}^2}{S(\mathbf{X})\sqrt{1/n_1 + 1/n_2}} \underset{H_0}{\sim} \mathcal{T}(n_1 + n_2 - 2)$$

http://fr.wikipedia.org/wiki/Loi_de_Student

Les arbres des deux parcelles ont-ils des tailles différentes ?

Type	Moyenne	Ecart-type	nb obs
1	25.66	1.24	11
2	24.64	1.43	10

- On reprend le modèle gaussien pour tester l'égalité des tailles des arbres des deux parcelles: $H_0 : \{\mu_1 = \mu_2\}$
- On fait l'hypothèse que les variances sont homogènes. A partir des données on obtient $S^2_{\text{pooled}}(\mathbf{x}) = 1.78$ (1.33 pour l'écart-type).
- La statistique de Student vaut donc $T(\mathbf{x}) = 1.75$
- La p-value du test $\mathbb{P}\{|T(\mathbf{X})| > 1.75\} = 0.096$ en considérant une $\mathcal{T}(10 + 11 - 2)$

Le test de Student à variances hétérogènes

- Si les deux groupes ont des variances différentes, **le modèle change**:

$$X_i^1 \sim \mathcal{N}(\mu_1, \sigma_1^2), \qquad X_i^2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$$

- Les variances ne peuvent plus être regroupées, mais sont estimées séparément

$$S^2(\mathbf{X}^1) = \overline{X^2}^1 - (\overline{X}^1)^2, \qquad S^2(\mathbf{X}^2) = \overline{X^2}^2 - (\overline{X}^2)^2$$

- Donc la statistique change également:

$$T(\mathbf{X}) = \frac{\overline{X}^1 - \overline{X}^2}{\sqrt{S^2(\mathbf{X}^1)/n_1 + S^2(\mathbf{X}^2)/n_2}}$$

- La loi de cette statistique est toujours une loi de Student sous H_0 mais ses degrés de liberté sont calculés par une approximation (Satterswaite)

Conclusion sur le test de Student: il faut d'abord tester l'égalité des variances !

Obs On considère des échantillons de **2 populations indépendantes** de taille n_1 et n_2, notés $\mathbf{x}^1 = (x_1^1, \ldots, x_{n_1}^1)$ et $\mathbf{x}^2 = (x_1^2, \ldots, x_{n_2}^2)$.

F_θ On suppose que la variable d'intérêt peut être modélisée par une **loi gaussienne**, telle que:

$$X_i^1 \underset{iid}{\sim} \mathcal{N}(\mu_1, \sigma_1^2), \; X_i^2 \underset{iid}{\sim} \mathcal{N}(\mu_2, \sigma_2^2)$$

H_0 On souhaite tester l'hypothèse **sur le paramètre de variance**

$$H_0 : \{\sigma_1^2 = \sigma_2^2\}, \qquad H_1 : \{\sigma_1^2 \neq \sigma_2^2\}$$

Test d'égalité des variances

- On estime les variances du modèle:

$$S^2(\mathbf{X}^1) = \overline{X^2}^1 - (\overline{X}^1)^2, \qquad S^2(\mathbf{X}^2) = \overline{X^2}^2 - (\overline{X}^2)^2$$

- On connait la loi des estimateurs sous H_0 (admis):

$$S_1^2(\mathbf{X}) \sim \sigma^2\chi^2(n_1 - 1), \qquad S_2^2(\mathbf{X}) \sim \sigma^2\chi^2(n_2 - 1).$$

- On construit la statistique de test de Fisher (ratio de deux χ^2 indépendants):

$$T(\mathbf{X}) = \frac{S_1^2(\mathbf{X})}{S_2^2(\mathbf{X})} \underset{H_0}{\sim} \mathcal{F}(n_1 - 1, n_2 - 1)$$

- Exemple avec les tailles d'arbres $T(\mathbf{x}) = 0.7511$, $\mathbb{P}(T(\mathbf{x}) > 0.7511) = 0.6593$.

Le test de Student sur données appariées

- L'hypothèse importante du test de Student est de savoir si les deux populations que l'on compare sont indépendantes ou non
- Cette hypothèse permet de calculer la variance des deux populations (somme pondérées de variances)
- Elle n'est pourtant pas toujours vérifiée. Exemple: données familiales (trio), mesures répétées sur le même individu
- Une première idée pour prendre en compte ces dépendances est de travailler sur la moyenne des différences

Outline

Tests asymptotiquement Gaussiens

- Si on dispose de beaucoup d'observations, alors le TCL nous aide à trouver la loi d'une statistique fondée sur $\overline{X}$!
- Exemple dans le cas binomial: $H_0 : \{p = p_0\}$, $\widehat{p}(\mathbf{X}) = \overline{X}$
- On utilise la statistique:

$$T(\mathbf{X}) = \frac{\widehat{p}(\mathbf{X}) - p_0}{\sqrt{p_0(1-p_0)}} \times \sqrt{n} \underset{H_0}{\sim} \mathcal{N}(0,1)$$

- Donc on utilisera la fonction de répartition Φ pour calculer la p-value
- Ce test asymptotique peut être utilisé dès que $np \geq 5$!

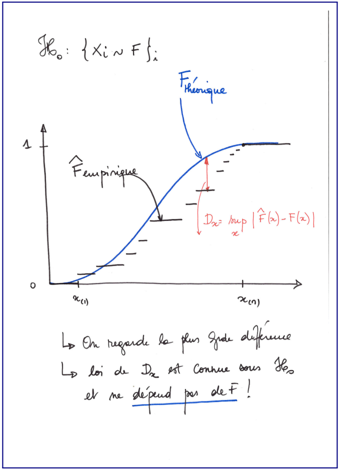
Outline

Avant d'affirmer que ...

- ...l'échantillon observé est distribué suivant une certaine loi F_θ, on peut s'interroger sur la pertinence de ce choix
- Les **tests d'ajustement** permettent:
  1. De déterminer si un échantillon est distribué comme une loi de référence (Gaussienne, Exponentielle...)
  2. De comparer les distributions de deux échantillons
- Comparer deux échantillons peut "simplement" consister à dire "sont-ils distribués différemment ou non ?" sans se focaliser sur la caractéristique qui est la cause de la différence
- Deux stratégies : le test de Kolmogorov Smirnov ou le test du χ^2 d'ajustement.

Le test de Kolmogorov Smirnov

- C'est un test usuel utilisé pour comparer des distributions
- On observe **x** et on souhaite savoir si l'hypothèse $X_i \sim \mathcal{N}(\mu, \sigma^2)$ est raisonnable
- On calcule la distance entre la fonction de répartition empirique $\widehat{F}$ et la fonction de répartion supposée sous $\mathcal{H}_0$
- La statistique de test considérée est la plus grande des différences entre $\widehat{F}$ et F_{theo}.



Introduction au test d'ajustement du χ^2

- Ce test permet de comparer la distribution empirique de comptages par rapport à une loi multinomiale
- Il permet de comparer soit une distribution empirique à une distribution théorique, ou deux distributions empiriques entre elles.
- Exemple avec le balanin de la châtaigne *Curculio elephas*. On a compté le nombre de parasites dans chacun des chataignes récoltées:
- Peut-on dire que le nombre de balanins suit une loi de Poisson de paramètre $\lambda = 0.3453$?

nb parasites	0	1	2	3	4	5	6	7	8	9	10	11
nb chataignes	1043	172	78	15	10	7	2	1	0	0	0	1

Le modèle multinomial pour le test du χ^2 d'ajustement

- Dans un premier temps, on considère la **loi Multinomiale**
- On note $\mathbf{N} = (N_1, \dots, N_k)$ un vecteur de comptages aléatoires t.q. $\sum_k N_k = n$.
- On dit que $\mathbf{N} \sim \mathcal{M}(n, p_1, \dots, p_k)$ si, pour $\sum_{j=1}^k p_j = 1$, et $\sum_{j=1}^k n_j = n$

$$\mathbb{P}\{N_1 = n_1, \dots, N_k = n_k\} = \frac{n! p_1^{n_1} \dots, p_k^{n_k}}{n_1! \dots n_k!},$$

- Cette loi modélise n tirages consécutifs avec remise dans une urne à k catégories n_1 boules de type 1, ... n_k boules de type k, avec des boules de type j en proportion p_j dans l'urne.
- Du fait de la contrainte $\sum_{j=1}^k n_j = n$, les comptages de différentes catégories ne sont pas indépendants !

$$\text{cov}\left(N_j, N_{j'}\right) = -p_j p_{j'}$$

Vers le test du χ^2 d'ajustement

- On observe $\mathbf{n} = (n_1, ..., n_k)$, et on souhaiterait savoir si $\mathbf{n}$ suit une loi multinomiale de paramètres $p_1, \ldots, p_k$
- L'hypothèse nulle dans le cas d'un test d'ajustement sera:

$$H_0 : \{\mathbf{N} = (N_1, \ldots, N_k)) \sim \mathcal{M}(n, p_1, \ldots, p_k)\}$$

- Dans le cas des balanins, on a 12 catégories (de 0 à 11),

$$\mathbf{n} = (1043, 172, 78, 15, 10, 7, 2, 1, 0, 0, 0, 1).$$

- Les proportions théoriques sont donnés par la loi de Poisson:

$$p_k = \frac{e^{-\lambda}\lambda^k}{k!}$$

Application aux balanins

- Certaines catégories ont des effectifs théoriques très faibles.

nb parasites	0	1	2	3	4	5	6	7	8	9	10	11
nb chataignes	1043	172	78	15	10	7	2	1	0	0	0	1
compt. attendu	940.94	324.91	56.10	6.46	0.56	0.04	0.0	0.0	0.0	0.0	0.0	0.0

- Dans ce cas, une stratégie consiste à regrouper des cases pour que les effectifs théoriques soient tous plus grands que 5.

nb parasites	0	1	2	$\geq$3
nb chataignes	1043	172	78	36
compt. attendu	940.94	324.91	56.10	7,14

Loi d'un vecteur multinomial

- On ne peut pas tester séparemment chaque catégorie car les comptages sont liés
- Grâce au TLC, on sait que

$$\begin{bmatrix} \overline{N}_1 - np_1 \\ \vdots \\ \overline{N}_{k-1} - np_{k-1} \end{bmatrix} \underset{H_0}{\sim} \mathcal{N}\left(\begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}, V \right)$$

- La statistique du χ^2 utilisée pour tester H_0 est définie par (admis)

$$T(\mathbf{N}) = \sum_{j=1}^{k-1} \frac{\left(\overline{N}_j - np_j\right)^2}{np_j} \underset{n\to\infty}{\sim} \chi^2(k-1)$$

Application aux balanins

- Certaines catégories ont des effectifs théoriques très faibles.

nb parasites	0	1	2	3	4	5	6	7	8	9	10	11
nb chataignes	1043	172	78	15	10	7	2	1	0	0	0	1
compt. attendu	940.94	324.91	56.10	6.46	0.56	0.04	0.0	0.0	0.0	0.0	0.0	0.0

- Dans ce cas, une stratégie consiste à regrouper des cases pour que les effectifs théoriques soient tous plus grands que 5.

nb parasites	0	1	2	$\geq$3
nb chataignes	1043	172	78	36
compt. attendu	940.94	324.91	56.10	7,14

- La statistique de test vaut $T(\mathbf{n}) = 208,26$,

$$\mathbb{P}\{T(\mathbf{N}) \geq 208,26\} = 6.610^{-27}, (\text{loi du } \chi^2(3))$$

Conclusion sur le test d'ajustement du χ^2

- Ce test permet de comparer la distribution empirique de comptages par rapport à une loi multinomiale
- La statistique utilisée pour tester l'hypothèse est une statistique dont la loi asymptotique est une loi du χ^2
- On peut également utiliser ce test pour comparer deux distributions
- Si X est une variable continue, on choisit un entier k et une suite croissante $t_1 < \ldots < t_{k-1}$ de telle manière que $X(\Omega)$ est découpé en k classes:

$$X \sim \mathbb{P}_0,\ p_j = \mathbb{P}_0\{t_{j-1} < X < t_j\}$$

- On fait ensuite le test en prenant les p_j ainsi construits comme référence sous H_0

Ajustement et test d'indépendance

- Une utilisation très courante du test du χ^2 est le test d'indépendance entre deux variables X et Y.
- On note p_i la loi marginale de X et q_j la loi marginale de Y (I et J valeurs resp):

$$p_i = \mathbb{P}(X = i), \qquad q_j = \mathbb{P}(Y = j)$$

- La loi du couple est donnée par:

$$p_{ij} = \mathbb{P}(X = i, Y = j)$$

- L'hypothèse d'indépendance se pose:

$$H_0 : \ p_{ij} = p_i \times q_j.$$

Construction du test d'indépendance

- On créée une variable N_{ij} qui compte combien de fois le couple (X, Y) a pris les valeurs (i, j) après n observations au total
- On construit une table de contingence qui résume l'information contenue dans les comptages:

	y_1	y_j	y_J	marge
x_1	N_{11}	N_{1j}	N_{1J}	N_{1+}
x_i	N_{i1}	N_{ij}	N_{iJ}	N_{i+}
x_I	N_{I1}	N_{Ij}	N_{IJ}	N_{I+}
marge	N_{+1}	N_{+j}	N_{+J}	N_{++}

- On estime les paramètres du modèle par les fréquences empiriques:

$$\widehat{p}_i = N_{i+}/N_{++}, \qquad \widehat{q}_j = N_{+j}/N_{++}, \qquad , \widehat{p}_{ij} = N_{ij}/N_{++}$$

Construction du test d'indépendance

- Les **marges** permettent de calculer l'effectif espéré de chaque case sous l'hypothèse d'indépendance

$$N_{++}\widehat{p}_i\widehat{q}_j = \frac{N_{i+} \times N_{+j}}{N_{++}}$$

- On calcule la statistique du χ^2 sous l'hypothèse nulle:

$$T(\mathbf{X}, \mathbf{Y}) = \sum_{i,j} \frac{(N_{ij} - n\widehat{p}_i\widehat{q}_j)^2}{n\widehat{p}_i\widehat{q}_j} \underset{n\to\infty}{\sim} \chi^2([I-1]\times[J-1])$$

Exemple des balanins

- Lors de la dissection des chataignes on a aussi découvert un autre insecte parasite, le carpocapse *Cydia splendana*.
- La distribution conjointe des deux parasites dans les 1329 chataignes est:

nb parasites	0	1	2	3	4	≥ 5	total
Abs de carpo	804	139	71	13	9	10	1046
Presence d'au moins 1	239	33	7	2	1	1	283
total	1043	172	78	15	10	11	1329

- On s'interroge donc sur la dépendance de la distribution des carpocaspes en fonction de celle des balanins.

Construction du test

- On construit le tableau des effectifs attendus:

nb parasites	0	1	2	3	4	$\geq$5	total
Abs de carpo	820,90	135,37	61,39	11,81	7,87	8,66	1046
Presence d'au moins 1	222,10	36,63	16,61	3,19	2,13	2,34	283
total	1043	172	78	15	10	11	1329

- Pour atteindre des effectifs attendus ≥ 5 on regroupe des cases:

Effectifs observés				
nb parasites	0	1	2	≥ 3
Abs de carpo	804	139	71	32
Presence d'au moins 1	239	33	7	4

Effectifs attendus				
nb parasites	0	1	2	≥ 3
Abs de carpo	820,90	135,37	61,39	28,34
Presence d'au moins 1	222,10	36,63	16,61	7,66

- La réalisation de la statistique de test $T(\mathbf{x}, \mathbf{y}) = 11.38$ que l'on compare au quantile de la loi du χ^2 à (4-1)(2-1) degrés de libertés (7.85 à $\alpha = 0.05$).

Outline

L'utilité des rangs

- Lorsque l'on souhaite étudier le lien entre deux variables d'unités différentes, ou que l'on ne souhaite pas faire d'hypothèse de distribution sur les observations
- Définition: si $(x_1, \ldots, x_n)$, on appelle **statistique d'ordre** $(x_{(1)}, \ldots, x_{(n)})$ t.q. $\widetilde{\mathbf{x}} = (x_{(1)} < \ldots < x_{(n)})$ et **vecteur des rangs** toute permutation R_x sur $\{1, \ldots, n\}$ t.q.:

$$\widetilde{x}_i = x_{R_x(i)}$$

- On utiliser le rho de Spearman pour quantifier la laison entre deux échantillons **x** et **y**:

$$\rho = 1 - \frac{6 \times \sum_i (R_x(i) - R_y(i))^2}{n(n^2 - 1)}$$

La télévision rend-elle intelligent ?

- On pose X_i le QI de l'individu i et Y_i le nombre d'heures passées par semaine devant la télévision.
- On souhaite quantifier le lien entre X et Y.
- On note $d_i = (R_x(i) - R_y(i))$ et on obtient $\rho = -0.175$

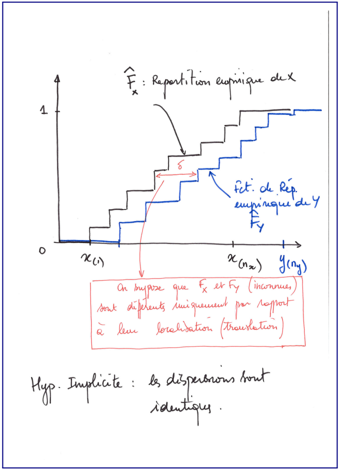
QI	h-TV.semaine	rang QI	rand h	d	d^2
86	0	1	1	0	0
97	20	2	6	-4	16
99	28	3	8	-5	25
100	27	4	7	-3	9
101	50	5	10	-5	25
103	29	6	9	-3	9
106	7	7	3	4	16
110	17	8	5	3	9
112	6	9	2	7	49
113	12	10	4	6	36

Contexte, introduction des tests de rang

- Lorsqu'il n'est pas raisonnable d'utiliser des lois limites (domaine de validité du TCL).
- Lorsque trop peu d'observations sont disponibles pour faire des hypothèses fortes quant à leur distribution
- Il existe une théorie générale fondée sur les **statistiques de rang** ("Rank Tests"-Hajek & Sidak, 1967)
- Le plus connu est le test de Wilcoxon /Mann-Whitney. Il existe aussi des tests de permutation.

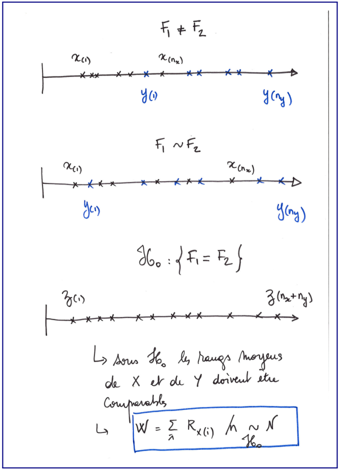
Modèle de localisation

- On considère F_X et F_Y deux fonctions de répartition t.q. $(X_i) \sim F_X$ et $(Y_j) \sim F_Y$.
- On suppose que F_X et F_Y ne diffèrent que par rapport à leur paramètre de localisation: $F_X(x) = H(x - \mu_X)$ et $F_Y(x) = H(x - \mu_Y)$
- On pose l'hypothèse nulle: $H_0 : \{\mu_X = \mu_Y\}$: **les deux lois ont la même localisation**
- Attention: les tests non paramétriques reposent aussi sur des hypothèses



Intuition du test de rang

- Test de Wilcoxon de comparaison de distributions: $H_0 : \{F_X = F_Y\}$ vs. $H_1 : \{X_i \sim F_X,\ Y_j \sim F_Y,\ \text{avec } F_X \leq F_Y\}$
- Idée: sous H_0 on mélange les deux échantillons: le nombre de couples (X_i, Y_j) avec $(X_i \leq Y_j)$ devrait être le même que le nombre de couples (X_i, Y_j) avec $(X_i \geq Y_j)$



Intuition du test de rang

- On considère Z, la statistique d'ordre de (X, Y) mélangés. Sous H_0 les rangs des X_i dans Z doivent se comporter de manière similaire aux rangs de Y dans Z.
- **Une** statistique des rangs des X dans Z est $W = \sum_i R_X(i)/(n_X + n_Y)$
- Sous H_0, W est approximativement Gaussien:

$$W \underset{H_0}{\sim} \mathcal{N}\left(\frac{n_X(n_X + n_Y + 1)}{2}, \frac{n_X n_Y(n_Y + n_X + 1)}{12}\right)$$