Introduction au statistique non-paramétrique

Hamel Elhadj

Departement de Mathematiques

Université Hassiba Benbouali-Chlef

Ce cours est distiné aux étudiants Master2 mathématiques

Option :Mathématique Apliquées et statistique

2021/2022

21 octobre 2021

Plan de la présentation

Plan de cours

1. **Introduction .**
2. **Chapitre 1 :** Les tests non paramétriques
3. **Chapitre 2 :** Estimation de fonction de répartition et de la densité
4. **Chapitre 3 :** Estimation non-paramétrique de fonction de régression
5. **Chapitre 4 :** Estimation non paramétrique des autre fonction (survie, hasard,....

Plan de cours

1. **Introduction .**
2. **Chapitre 1 :** Les tests non paramétriques
3. **Chapitre 2 :** Estimation de fonction de répartition et de la densité
4. **Chapitre 3 :** Estimation non-paramétrique de fonction de régression
5. **Chapitre 4 :** Estimation non paramétrique des autre fonction (survie, hasard,....

Plan de cours

1. **Introduction .**
2. **Chapitre 1 :** Les tests non paramétriques
3. **Chapitre 2 :** Estimation de fonction de répartition et de la densité
4. **Chapitre 3 :** Estimation non-paramétrique de fonction de régression
5. **Chapitre 4 :** Estimation non paramétrique des autre fonction (survie, hasard,....

Plan de cours

1. **Introduction .**
2. **Chapitre 1 :** Les tests non paramétriques
3. **Chapitre 2 :** Estimation de fonction de répartition et de la densité
4. **Chapitre 3 :** Estimation non-paramétrique de fonction de régression
5. **Chapitre 4 :** Estimation non paramétrique des autre fonction (survie, hasard,....

Plan de cours

1. **Introduction .**
2. **Chapitre 1 :** Les tests non paramétriques
3. **Chapitre 2 :** Estimation de fonction de répartition et de la densité
4. **Chapitre 3 :** Estimation non-paramétrique de fonction de régression
5. **Chapitre 4 :** Estimation non paramétrique des autre fonction (survie, hasard,....

Introduction

Introduction.

L'estimation statistique est un domaine très important de la statistique mathématique qui développe des techniques pour décrire certaines caractéristiques d'ensembles d'observations. On distingue deux composantes principales, à savoir, l'estimation paramétrique et l'estimation non paramétrique.

Statistique non paramétrique : c'est quoi ?

La statistique paramétrique est le cadre "classique" de la statistique.

Le modèle statistique y est décrit par un nombre fini de paramètres.

Typiquement $\mathcal{M} = \{\mathbb{P}_\theta; \theta \in \mathbb{R}\}$ est le modèle statistique qui décrit la distribution des variables aléatoires observées.

Exemples

- Observations réelles avec un seul mode :
  modèle Gaussien. $\mathcal{M} = \{\mathcal{N}(\mu, \sigma^2); \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}^{*+}\}$
- Observations réelles avec plusieurs modes : modèle de mélange Gaussien.
- Observations de comptage : modèle loi Poisson. $\mathcal{M} = \{\mathbb{P}(\lambda); \lambda \in \mathbb{R}^{*+}\}$

Statistique non paramétrique : c'est quoi ?

Par opposition, en statistique non paramétrique, le modéle n'est pas décrit par un nombre fini de paramètres. Divers cas de figures peuvent se présenter, comme par exemple :

- On s'autorise toutes les distributions possibles, i.e. on ne fait aucune hypothèse sur la forme/nature/type de la distribution des variables aléatoires.
- On travaille sur des espaces fonctionnels, de dimension infinie.
  Exemple : les densités continues sur $[0, 1]$, ou les densités monotones sur $\mathbb{R}$
- Le nombre de paramètres du modèle n'est pas fixé et varie (augmente) avec le nombre d'observations.
- Le support de la distribution est discret et varie (augmente) avec le nombre d'observations.

Statistique non paramétrique : c'est quoi ?

Définition

Un modèle est dit **non-paramétrique** *lorsque qu'il ne peut se mettre sous forme paramétrique (c'est à dire que Y observé est défini par un paramètre $\theta \in \Theta \in \mathbb{R}^k, \quad k \in N*$).*

On cherche à estimer une fonction appartenant à un espace fonctionnel (infini ou de grande dimension $D = D_n \longrightarrow \infty$).

À partir de n observations recueillies par sondage il pose :

$Y_i = f(X_i) + "\varepsilon_i, \ "\varepsilon_i \sim N(0; \sigma^2) i.i.d.$

- le modèle est linéaire si $f(X_i) = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2}$
- et non-paramétrique si f est quelconque.

Quand est-ce qu'on l'utilise ?

1. Quand on n'arrive pas à ajuster correctement les observations avec une distribution paramétrique.
2. Quand on n'a aucune idée de modèle, ou qu'on ne veut pas avoir un a priori sur le modèle.
3. Quand on ne sait pas combien de composantes on veut mettre dans un mélange.
4. Quand le nombre de variables est trop grand (problème de grande dimension), trop de paramètres.
5. ...

Quand est-ce qu'on l'utilise ?

1. Quand on n'arrive pas à ajuster correctement les observations avec une distribution paramétrique.
2. Quand on n'a aucune idée de modèle, ou qu'on ne veut pas avoir un a priori sur le modèle.
3. Quand on ne sait pas combien de composantes on veut mettre dans un mélange.
4. Quand le nombre de variables est trop grand (problème de grande dimension), trop de paramètres.
5. ...

Quand est-ce qu'on l'utilise ?

1. Quand on n'arrive pas à ajuster correctement les observations avec une distribution paramétrique.
2. Quand on n'a aucune idée de modèle, ou qu'on ne veut pas avoir un a priori sur le modèle.
3. Quand on ne sait pas combien de composantes on veut mettre dans un mélange.
4. Quand le nombre de variables est trop grand (problème de grande dimension), trop de paramètres.
5. ...

Quand est-ce qu'on l'utilise ?

1. Quand on n'arrive pas à ajuster correctement les observations avec une distribution paramétrique.
2. Quand on n'a aucune idée de modèle, ou qu'on ne veut pas avoir un a priori sur le modèle.
3. Quand on ne sait pas combien de composantes on veut mettre dans un mélange.
4. Quand le nombre de variables est trop grand (problème de grande dimension), trop de paramètres.
5. ...

Quand est-ce qu'on l'utilise ?

1. Quand on n'arrive pas à ajuster correctement les observations avec une distribution paramétrique.
2. Quand on n'a aucune idée de modèle, ou qu'on ne veut pas avoir un a priori sur le modèle.
3. Quand on ne sait pas combien de composantes on veut mettre dans un mélange.
4. Quand le nombre de variables est trop grand (problème de grande dimension), trop de paramètres.
5. ...

Statistique non paramétrique

- Avantages
  - Moins d'a priori sur les observations,
  - Modèles plus généraux, donc plus robustes au modéle.
- Inconvénients
  - Les tests paramétriques, quand leurs conditions sont remplies, sont plus puissants que les tests non-paramétriques.
  - Vitesses de convergence plus lentes = il faut plus de données pour obtenir une précision équivalente.

Quelques références bibliographiques pour ce cours

- L. Wasserman. *All of nonparametric statistics.* Springer, 2006.
- E.L. Lehmann. *Elements of large sample theory.* Springer Texts in Statistics. Springer, 1999.
- A. B. Tsybakov . *Introduction à l'estimation non-paramétrique,* Springer, 2004.
- D. Bosq. *Nonparametric statistics for stochastic processes,*Springer, 1996
- F. Comte . *Estimation non-paramétrique.* (2015) Spartacus IDH
- E. Giné & R. Nickl. *Mathematical foundations of infinite-dimensional statistical models.* (2015) Cambridge University Press

Chap01 : Les tests non paramétriques

Introduction

Un test d'hypothèse consiste à choisir entre deux hypothèses incompatibles en se fondant sur des résultats d'échantillonnage.

L'une des deux hypothèses à tester est généralement privilégiée par rapport à l'autre : on tient à limiter à priori la probabilité de la rejeter àtort. Cette hypothèse désigne traditionnellement les situations d'absence de changement par rapport à un statuquo, ouencore l'absence de différence entre des paramètres.

Cette hypothèse, notée H_0 , est appelée *hypothèse nulle.*

L'autre hypothèse, notée H_1, est appelée *hypothèse alternative.*

Deux familles de tests

- **Tests paramétriques :** tests d'hypothèses relatives à un ou plusieurs paramètres d'une ou plusieurs variables aléatoires de *lois connues.*
- **Tests non paramétriques :**(Distribution-free tests) tests ne nécessitant pas d'hypothèses sur la distribution sous-jacente.

La conclusion d'un test d'hypothèse se fait en terme de rejet ou de non-rejet de l'hypothèse nulle.

Pour des petits échantillons, on est utilise plutôt des tests non paramétriques, sauf si la variable étudiée suit une loi normale.

Définition

Les tests sont non paramétriques lorsque la distribution des variables aléatoires n'est pas spécifiée sous au moins une des deux hypothèses (nulle ou alternative).

Exemple : Tests d'adéquation à une loi , Tests de comparaison ,Tests d'indépendance.

Tests paramétrique et non paramétrique

La plupart des tests statistiques sont construits à partir d'hypothèses sur les distributions des variables étudiées chez les individus. Dans un grand nombre de situations, la distribution utilisée est la loi normale.

L'utilisation d'un test paramétrique suppose de *connaître la loi (ou la famille de lois)* sous-jacente et que les densités de probabilités associées dépendent de *paramètres donnés de la loi tels la moyenne et la variance pour la loi normale* .

Lorsque la famille à laquelle appartiennent les densités de probabilités *est inconnue*, on optera pour un test non paramétrique.

Tests non-paramétrique

Lorsqu'on ne peut pas supposer que les variables sont normales et de même variance, on peut utiliser des *tests dits non paramétriques* qui sont valables quelles que soient les lois des variables de base.

Avantages des tests non paramétriques

1. Leur emploi se justifie lorsque les conditions d'applications des autres méthodes ne sont pas satisfaites, même après d'éventuelles **transformations** de variables.
2. Pour des échantillons de taille trés faible , la seule possibilité est l'utilisation d'un test non paramétrique sauf si la distribution de la population est connue on utilise un test paramétrique

Tests non-paramétrique

Lorsqu'on ne peut pas supposer que les variables sont normales et de même variance, on peut utiliser des *tests dits non paramétriques* qui sont valables quelles que soient les lois des variables de base.

Avantages des tests non paramétriques

1. Leur emploi se justifie lorsque les conditions d'applications des autres méthodes ne sont pas satisfaites, même après d'éventuelles **transformations** de variables.
2. Pour des échantillons de taille trés faible , la seule possibilité est l'utilisation d'un test non paramétrique sauf si la distribution de la population est connue on utilise un test paramétrique

Test de signe - Test de médiane (ou de symétrie)

Le test du signe est une procédure non paramétrique relativement simple pour tester des hypothèses sur la tendance centrale (la médiane) d'une distribution de probabilité non normale .

Notez que nous avons utilisé la médiane plutôt que la moyenne de la population. C'est parce que le test de signe, comme de nombreuses procédures non paramétriques, fournit des inférences sur la médiane plutôt que la moyenne de la population .et, comme est moins affecté par l'asymétrie de la distribution et la présence de valeurs aberrantes (observations extrêmes). (moins sensible aux valeurs extrêmes)

Test de signe - Test de médiane (ou de symétrie)

Ce test .

- s'appelle aussi Tests de médiane (ou de symétrie)
- est un cas particulier d'un test plus général, appelé le test binomiale
- **Objectif général.** Soient $X_1, ..., X_n$ v.a. réelles i.i.d. On veut tester si la distribution de X est symétrique par rapport à 0.

On observe un échantillon $X_1, ..., X_n$ de v.a. réelles i.i.d. On teste l'hypothèse :

- $H_0 : P(X \leq 0) = 1/2$ i.e. "la médiane de la distribution est nulle" contre
- $H_1^+ : P(X \leq 0) > 1/2$ i.e. "la médiane de la distribution est négative" ou
- $H_1^- : P(X \leq 0) < 1/2$ i.e. "la médiane de la distribution est positive"

Tests de médiane

où bien Les hypothèses considérées sont de la forme :

- H_0 : la médiane de la distribution est égale à M_0, contre
- H_1 : la médiane de la distribution est différente de M_0 où M_0 est *une valeur constante fixée a priori.*
- Statistique de signe $\sum_{i=1}^{n} \mathbb{I}_{X_i < M_0 = 0}$
  où $p = \mathbb{P}(X \leq M_0)$. Sous $H_0 : p = 1/2$, on a $S_+ \sim\sim \mathcal{B}in(n, 1/2)$ et sous $H_1 : P(X < M_0) > 1/2$, la statistique S_+ est stochastiquement plus grande que sous H_0. On rejette donc H_0 pour les grandes valeurs de S_+.

Principe du test

Le test du signe consiste à remplacer les observations plus grandes ou égales à M_0 par un signe "+" et celles qui lui sont inférieures par un signe "-".(S^+ , S^-)

les etapes de test de signe :

- Comptez le nombre des x_i qui dépassent M_0. Appelez S_+ (les petits S_-).
- Rejeter H_0 si S_+ est trop grand (ou si S_- est trop petit).

Quelle doit être la taille de S_+ pour rejeter ? Pour le savoir, nous devons connaître la distribution du v.a pour S_+.

soit $p = \mathbb{P}(X_i > M_0)$ et $1 - p = \mathbb{P}(X_i < M_0)$.

S^+ est un somme de Bernoulli. Alors c'est un binômial ! ! ! !

$S^+ \sim \mathcal{B}in(n, p)$ et $S^- \sim \mathcal{B}in(n, 1 - p)$.

Principe du test

Maintenant, si H_0 est vrai, M_0 est la vraie médiane et $p = 1/2$, donc :

$S^+ \sim \mathcal{B}in(n, 1/2)$ et $S^- \sim \mathcal{B}in(n, 1/2)$. .

Donc on rejeter H_0 si $S^+ \geq b_{n,\alpha}$ où $b_{n,\alpha}$, est le point critique α supérieur pour $\mathcal{B}in(n, 1/2)$.

(où rejetter quand $S^- \leq b_{n,1-\alpha}$)

Calculons maintenant la $p - valeur$ $(p - value)$ en utilisant la distribution binomiale :

$$\begin{aligned} p - value &= \mathbb{P}(S^+ \geq s^+) = \textstyle\sum_{i=s^+}^{n} \mathbb{C}_n^i \left(\frac{1}{2}\right)^n \\ &= \mathbb{P}(S^- \leq s^-) = \textstyle\sum_{i=0}^{s^-} \mathbb{C}_n^i \left(\frac{1}{2}\right)^n \end{aligned}$$

Test de signe

Propriété

- *Pour les petites valeurs de n, la distribution $\mathcal{B}in(n, 1/2)$ est tabulée. Pour les grandes valeurs de n, on a recours à une approximation Gaussienne.*
- *Ce test est très général, mais il utilise très peu d'information sur les variables (uniquement leur signe, pas leurs valeurs relatives). C'est donc un test peu puissant.*
- *Le test de signe et rang utilise plus d'information sur les variables.*
- *Remarque : c'est en fait un test paramétrique ! puisque la loi de S_n sous H_0 et sous l'alternative est paramétrique $(\mathcal{B}in(n, p))$.*

exemple (test de signe)

Example:

The following data constitute a random sample of 15 measurement of the octane rating of a certain kind gasoline:

99.0 102.3 99.8 100.5 99.7 96.2 99.1 102.5 103.3 97.4 100.4
98.9 98.3 98.0 101.6

Test the null hypothesis $M = 98.0$ against the alternative hypothesis $M > 98.0$ at the 0.01 level of significance.

Solution:

$M_0 = 98.0$

99.0	102.3	99.8	100.5	99.7	96.2	99.1	102.5	103.3	97.4	100.4
+	+	+	+	+	–	+	+	+	–	+
				98.9	98.3	98.0	101.6			
				+	+	0	+			

Number of + sign, $S+ = 12$
Number of sample, n = 14 (15 -1) **why?**
$p = 0.5$ $=1/2$

exemple (test de signe)

1. H_0 : M ≤ 98.0

   H_1 : M > 98.0

2. $\alpha = 0.01$, Reject H_o if $p-\text{value} < 0.01$

3. From binomial probability table for S+ $= 12$, $n = 14$ and $p = 0.5$

   S+ $\sim b(14, 0.5)$, $p-\text{value} = P(\text{s} \geq 12) = 1 - P(\text{s} \leq 11) = 1 - 0.9935 = 0.0065$

4. Since $p-\text{value} = 0.0065 < 0.01 = \alpha$, thus we reject H_0 and accept H_1 and conclude that the median octane rating of the given kind of gasoline exceeds 98.0

Application :

exemple (test de signe)

Example: Suppose we have a random sample of bass from North Lake. The weights of the fish are 1.2, 4.3, 2.4, 1.4, 0.8, 1.9, 1.5, and 1.1 pounds. Suppose we want to test

H_0: $m = 2$ pounds versus H_a: $m < 2$ pounds,

where m is the *median* weight of bass in the lake.

if H_0 is true, the number of values above 2.0 is a binomial random variable with probability $p = .5$ and $n = 8$.

We compute the p-value as our data (six of eight bass less than two pounds) or more extreme (seven or eight of eight bass less than two pounds), under H_0 (median is two pounds). Remember that "more extreme" means "supports the alternative more." The p-value is the binomial probability of at least six out of eight, when $p = 0.5$. This is

$$\binom{8}{6}(.5)^6(.5)^2 + \binom{8}{7}(.5)^7(.5)^1 + \binom{8}{8}(.5)^8(.5)^0 = 0.1445.$$

There is not enough evidence to reject the null hypothesis at $\alpha = 0.10$. The interpretation of the p-value in terms of repeated samples is: If the median bass weight is two pounds, and we draw many random samples of size eight from the lake, about 14.45% of these samples will contain at least six fish less than two pounds.

Application :

exemple (test de signe)

Application sous R :

```
> eye<-c(1.2,4.3,2.4, 1.4,0.8, 1.9, 1.5,1.1)
> S<-sum(eye >= 2)
> S
[1] 2
> binom.test(S, length(eye), 0.5, alternative="less")

        Exact binomial test

data:  S and length(eye)
number of successes = 2, number of trials = 8, p-value = 0.1445
alternative hypothesis: true probability of success is less than 0.5
95 percent confidence interval:
 0.0000000 0.5996894
sample estimates:
probability of success
                  0.25
```

exemple (test de signe)

Example: Twelve measurements of soil contamination in parts per million of a certain chemical were taken at random coordinates within five miles of a production plant:

6.2, 4.1, 3.5, 5.1, 5.0, 3.6, 4.8, 4.1, 3.6, 4.7, 4.3, 4.2.

If the soil contamination in the region is above 4.0, the plant must take steps to clean the soil and curtail pollution in the future. We want to test

$$H_0\text{: median soil contamination is 4.0 ppm}$$

versus

$$H_a\text{: median soil contamination is more than 4.0 ppm}$$

at $\alpha = .05$. (The production plant operators would like to see a smaller α, and the local grassroots environmental group would like to see a larger α.)

The p-value is the probability of our data (nine of twelve above 4.0) or more extreme (more than nine out of twelve), under the null hypothesis (each of the twelve measurements has probability 0.5 of being more than 4.0). We can use the R command `1-pbinom(8,12,.5)` to get $p = .073$. We accept H_0 at $\alpha = .05$, but we would reject at $\alpha = .10$.

exemple (test de signe)

Application sous R :

```
 H<-c(6.2, 4.1, 3.5, 5.1, 5.0, 3.6, 4.8, 4.1, 3.6, 4.7, 4.3, 4.2)
 SS<-sum(H >= 4)
 SS
[1] 9
> binom.test(SS, length(H), 0.5, alternative="greater")

        Exact binomial test

data:  SS and length(H)
number of successes = 9, number of trials = 12, p-value = 0.073
alternative hypothesis: true probability of success is greater than 0.5
95 percent confidence interval:
 0.4726734 1.0000000
sample estimates:
probability of success
                  0.75
```

Test de signe et rang (ou Wilcoxon signed rank test)

Rappel : Statistiques d'ordre et de rang

Soient $X_1, ..., X_n$ v.a. réelles.

Définition

- *La statistique d'ordre :* $\{(X_{(1)}, ..., X_{(n)})\}$ *est obtenue par réarrangement croissant des* X_i.

  Ainsi : $X_{(1)} \leq X_{(2)}, ..., \leq X_{(n)}$
- *Les rangs de* X *: Le vecteur* R_X *des rangs de* X *une permutation de* $\{1, .., n\}$ *telle que* $X_i = X_{(R_{X_{(i)}})}$.

Test de signe et rang (ou Wilcoxon signed rank test)

Soit Soient $X_1, ..., X_n$ v.a. réelles. un échantillon de v.a. réelles de loi supposée diffuse((i.e. $\mathbb{P}[X = x] = 0$ pour tout x). On veut tester :

–H0 : "la loi de X est symétrique (par rapport à 0)" contre

–H1 : "la loi de X n'est pas symétrique".

- Statistique de Wilcoxon :

$$W^+ = \sum_{i=1}^{n} R_{|X|}(i)\mathbb{I}_{X_i>0}$$

  où $R_{|X|}$ le vecteur des rangs associé à $(|X_1|, ..., |X_n|)$

- soit

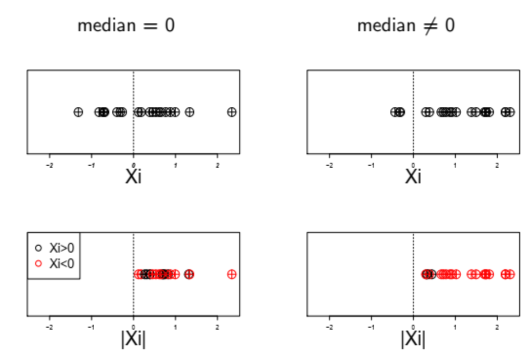
$$W^- = \sum_{i=1}^{n} R_{|X|}(i)\mathbb{I}_{X_i<0}$$

- On a $W^+ + W^- = \frac{n(n+1)}{2}$ p.s (vérifier) ? ? ?

Il existe des variantes prenant en compte des ex-aequos. ! ! !

Test de signe et rang (ou Wilcoxon signed rank test)

Exemple :



Lorsque la mediane de X est différente de 0, les rangs des X_i positifs ne sont pas distribués uniformément sur $\{1, ..., n\}$.

Test de signe et rang (ou Wilcoxon signed rank test)

thèoréme

Sous H_0 : "la loi de X est symétrique par rapport à 0", W_n^+ est libre en loi. De plus, :

- $\mathbb{E}[W_n^+] = \frac{n(n+1)}{4}$
- $V[W_n^+] = \frac{n(n+1)(2n+1)}{24}$
- $\frac{W_n^+ - \mathbb{E}[W_n^+]}{\sqrt{V[W_n^+]}} \rightsquigarrow^L \mathcal{N}(0,1)$

- Test de H_0. On rejette H_0 pour les grandes valeurs de $[W_n^+ - \frac{n(n+1)}{4}]$.
- La loi de W_n^+ sous H_0 est tabulée usuellement pour $n \leq 20$, et une approximation est calculée pour $n > 20$.

Test de signe et rang (ou Wilcoxon signed rank test)

Procédure de test :

1. Pour chacune des valeurs observées, trouvez la différence entre chaque valeur et la médiane $d_i = X_i - M_0$;
   où M_0 une valeur médiane qui a été spécifiée.
2. En ignorant l'observation où $d_i = 0$, classez les valeurs $|d_i|$ telque le plus petit aura un rang egale à 1. Si on a deux différences ou plus ont la même valeur, en utilise leurs rang moyen .
3. Pour l'observation $x_i > M_0$, indiquez le rang sous forme $R^+(d_i)$ et pour $x_i < M_0$ indiquez le rang sous forme $R_(d_i)$.
4. W^+ = la somme des rangs où la différence est positive , W^- = la sommes des rangs ou la difference soit négatives.

Exemple

Dans le tableau ci-dessous. En utilisant le test des rangs signés de Wilcoxon avec $\alpha = 0,05$, pouvons-nous conclure que l'eau potable de la communauté pourrait être égale ou supérieure à la limite recommandée de 40,0 ppm ? d'un échantillon de 11 ménages de la communauté

Household	Observed concentration (x)
A	39
B	20.2
C	40
D	32.2
E	30.5
F	26.5
G	42.1
H	45.6
I	**42.1**
J	29.9
K	40.9

Exemple

$m_0 = 40$

Household	Observed concentration x_i	$d_i = x_i - m_0$	$\lvert d_i \rvert$	Rank, $R(d_i)$	$+R(d_i)$	$-R(d_i)$
A	39	-1	1	2		2
B	20.2	-19.8	19.8	10		10
C	40	0	_	_		
D	32.2	-7.8	7.8	6		6
E	30.5	-9.5	9.5	7		7
F	26.5	-13.5	13.5	9		9
G	42.1	2.1	2.1	3.5	3.5	
H	45.6	5.6	5.6	5	5	
I	42.1	2.1	2.1	3.5	3.5	
J	29.9	-10.1	10.1	8		8
K	40.9	0.9	0.9	1	1	
				$\sum$	$T^+ = 13$	$T^- = 42$

Exemple

- H_0 : la mediane = 40 contre H_1 :la médiane < 40 , un échontillon , n=10 ! ! !
- sous H_1 $W^+ = T^+ = 13$
- À partir du tableau de Wilcoxon à un seul echontillon ,

  $\alpha = 0.05, \quad n = 10, \qquad \Longrightarrow w_{tab} = 10$

  on rejette H_0 si $W^+ < w_{tab}$
- conclusion : $W^+ = 13 > w_{tab} = 10$ On ne rejete pas H_0 et conclu que l'approvisionnement en eau de la ville pourrait contenir au moins 40,0 ppm .

résumé :

Case	H_0	H_1	Rejection region
Two tail	H_0: median $R(d) = m_0$	H_1 : median $R(d) \neq m_0$	$\min\left(T^+, T^-\right) \leq a$
Right tail	H_0 : median $R(d) = m_0$	H_1 : median $R(d) > m_0$	$T^- \leq a$
Left tail	H_0 : median $R(d) = m_0$	H_1 : median $R(d) < m_0$	$T^+ \leq a$

A Large-Sample Wilcoxon Signed-Rank Test for a Matched-Pairs Experiment: $n > 25$

Null hypothesis: H_0 : The population relative frequency distributions for the X's and Y's are identical.

Alternative hypothesis: (1) H_a : The two population relative frequency distributions differ in location (a two-tailed test),

or (2) the population relative frequency distribution for the X's is shifted to the right (or left) of the relative frequency distribution of the Ys (one-tailed tests).

Test statistic: $Z = \dfrac{T^+ - [n(n+1)/4]}{\sqrt{n(n+1)(2n+1)/24}}$.

Rejection region: Reject H_0 if $z \geq z_{\alpha/2}$ or $z \leq -z_{\alpha/2}$ for a two-tailed test. To detect a shift in the distributions of the X's to the right of the Y's, reject H_0 when $z \geq z_\alpha$. To detect a shift in the opposite direction, reject H_0 if $z \leq -z_\alpha$.

The Web site www.missingkids.com provides a searchable database of missing children. The ages of the following six children were obtained from this database.

Child	Adam	Juan	Benjamin	Samantha	Kayleen	Aiko
Age	4	9	5	7	6	3

Test, using level of significance $\alpha = 0.10$, whether the population median age of the missing children equals 6 years old. $M_0 = 6$.

Solution

$$H_0 : M = 6 \quad \text{versus} \quad H_a : M \neq 6$$

Step 1 **State the hypotheses.** We have a two-tailed test:

Step 2 We have a two-tailed test, with level of significance $\alpha = 0.10$ and $n = 5$, which gives us $T_{crit} = 1$. The rejection rule is to reject H_0 if $T_{data} \leq 1$.

Step 3

a. Find $d = \text{age} - M_0 = \text{age} - 6$ for each child

b. The absolute values of the differences $|d|$

c. We rank the absolute differences. T_{data} = the smaller of T_+ and $|T_-|$.

d. $T_+ = 4.5 + 1.5 = 6$. $|T_-| = |-9| = 9$ Thus, $T_{data} = 6$.

Step 4 **State the conclusion and the interpretation.** reject H_0 if $T_{data} \leq 1$. Because $T_{data} = 6$ is not ≤ 1, we do not reject H_0.

```
W<-c(39,20.2,40,32.2, 30.5,26.5,42.1, 45.6,42.1, 29.9,40.9)
  response<- W-40
  wilcox.test(response,alternative="greater",conf.int=TRUE)

        Wilcoxon signed rank test with continuity correction

data:  response
V = 13, p-value = 0.937
alternative hypothesis: true location is greater than 0
95 percent confidence interval:
 -10.09997       Inf
sample estimates:
(pseudo)median
     -4.399961
```

Table du test des signes et des rangs de Wilcoxon

la table ci-dessous (test unilatéral ou test bilatéral).

Test bilatéral			Tests unilatéraux		
n	risque 5%	risque 1%	n	risque 5%	risque 1%
6	0		6	2	
7	2		7	2	
8	3	0	8	5	
9	5	1	9	8	2
10	8	3	10	10	4
11	10	5	11	13	7
12	13	9	12	17	9
13	17	9	13	21	12
14	21	12	14	25	15
15	25	15	15	30	19
16	29	19	16	35	23
17	34	23	17	41	27
18	40	27	18	47	32

Tests sur deux populations

Introduction

soient $(X_1, ..., X_n)$ et $(Y_1, ..., Y_m)$ deux échantillons indépendants de v.a. réelles diffuses de lois respectives F et G. On veut tester H_0 : "$F = G$" contre H_1 : "$F \neq G$".

Pour des échantillons appariés

- $n = m$ et on se ramène à un unique échantillon $(D_1, ..., D_n)$ avec $D_i = X_i - Y_i$. (la difference entre les deux échantillons)
- Sous H_0, la loi de D est symétrique. D'où le test revient H_0' : "La loi de D est symétrique" contre H_0' :"La loi de D n'est pas symétrique".
- On applique le test de signe et et de rang de Wilcoxon.

Pour des échantillons non-appariés

- Test de Kolmogorov Smirnov de comparaison de 2 échantillons.
- Test de la somme des rangs de Wilcoxon (ou Mann-Whitney).

Introduction(Echantillon appariés/non appariés)

Soient $(X_1, ..., X_n)$ et $(X_1, ..., X_n)$ deux échantillons i.i.d. tirés dans deux populations de loi respectives F et G.

- Les échantillons sont dits appariés si X_i et Y_i sont associés dans le design de l'expérience :
  - X_i et Y_i sont des mesures d'une même quantité pour un même individu, par exemple à deux temps différents ou sous deux traitements différents.
  - X_i et Y_i sont deux quantités différentes mesurées sur un même individu.
- Si les deux échantillons sont tirés de façon indépendante dans deux populations, ils sont non apparièes.
- Si les échantillons sont appariés, ils sont nécessairement de même taille (n = m)

le test de signe et et de rang de Wilcoxon(The Wilcoxon Signed rank test for PAIRED SAMPLE)

procedure de test :

1. calcule la difference entre les deux echontillon : $d_i = X_i - Y_i$
2. En ignorant l'observation où $d_i = 0$, classez les valeurs $|d_i|$ telque le plus petit aura un rang egale à 1. Si on a deux différences ou plus ont la même valeur, en utilise leurs rang moyen .
3. T^+ : la somme des rangs où la difference est positive $X_i - Y_i > 0$, T^- : la somme des rang où la difference est négative $X_i - Y_i < 0$.
4. Comparez la statistique de test, $W = \min(T^+, T^-)$ avec la valeur critique dans les tableaux (voir tableau de Wilcoxon).

   L'hypothèse nulle est rejetée si. $W < w_{tab}$

le test de signe et et de rang de Wilcoxon(The Wilcoxon Signed rank test for PAIRED SAMPLE)

Test des rangs signés de Wilcoxon

Statistique de test

- T^+ grand par rapport à T^- $\rightarrow$ Les valeurs de X_1 sont stochastiquement plus élevées que celles de X_2
- T^- grand par rapport à T^+ $\rightarrow$ Les valeurs de X_1 sont stochastiquement plus faibles que celles de X_2

Si $\mathcal{H}_0$ est vraie alors $T^+ = T^- = \frac{1}{2}(\frac{n(n+1)}{2})$

Aussi, la statistique de test a pour expression

$$T = \min\left(T^-; T^+\right)$$

$$\mathbb{E}[T] = \frac{n(n+1)}{4} \qquad \mathbb{V}[T] = \frac{n(n+1)(2n+1)}{24}$$

Cas particulier si $n > 20$:

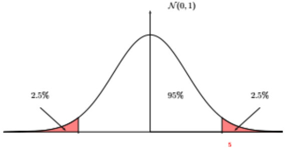
$$Z = \frac{T - \mathbb{E}[T]}{\sqrt{\mathbb{V}[T]}} \sim \mathcal{N}(0,1)$$

le test de signe et et de rang de Wilcoxon(The Wilcoxon Signed rank test for PAIRED SAMPLE)

Cas particulier si $n > 20$

$$Z = \frac{U - \frac{n_1 n_2}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \sim \mathcal{N}(0,1)$$

$$R.C. \; : \; |Z| \geq z_{1-\alpha/2}$$



Retour à l'exemple

Valeurs	1	1	1	1	2	2	2	4
Rangs bruts	1	2	3	4	5	6	7	8
Rangs finaux	2.5	2.5	2.5	2.5	6	6	6	8
Signes	+	+	-	+	-	-	+	+

$T^+ = 2.5 + 2.5 + 2.5 + 6 + 8 = 21.5$ $\qquad$ $T^- = 2.5 + 6 + 6 = 14.5$

$$T = \min\left(T^-; T^+\right) = 14.5$$

Lecture dans la table : $T_{lim} = 4 \rightarrow$ N.S.

Example

En utilisant le test des rangs signés de Wilcoxon avec $\alpha = 0,10$, pouvons-nous conclure que la différence médiane pour la population d'une telle tâche pourrait être nulle ?

Computing task	Time required for software packages	
	x_i	y_i
A	24	23.1
B	16.7	20.4
C	21.6	17.7
D	23.7	20.7
E	37.5	42.1
F	31.4	36.1
G	14.9	21.8
H	37.3	40.3
I	17.9	26
J	15.5	15.5
K	29	35.4
L	19.9	25.5

Example

Solution:

Computing task	Time required for software packages x_i	y_i	$d_i = x_i - y_i$	$\lvert d_i \rvert$	Rank, $R(d_i)$	$+R(d_i)$	$-R(d_i)$
A	24	23.1	0.9	0.9	1	1	
B	16.7	20.4	-3.7	3.7	4		4
C	21.6	17.7	3.9	3.9	5	5	
D	23.7	20.7	3	3	2.5	2.5	
E	37.5	42.1	-4.6	4.6	6		6
F	31.4	36.1	-4.7	4.7	7		7
G	14.9	21.8	-6.9	6.9	10		10
H	37.3	40.3	-3	3	2.5		2.5
I	17.9	26	-8.1	8.1	11		11
J	15.5	15.5	0	0	_		_
K	29	35.4	-6.4	6.4	9		9
L	19.9	25.5	-5.6	5.6	8		8
					$\sum$	$T^+ = 8.5$	$T^- = 57.5$

Exemple

1. H_0: median $= 0$

   H_1: median $\neq 0$ (two tail test)

2. Based on the alternative hypothesis, the test is $\min(T^+, T^-) = \min(8.5, 57.5) = 8.5$
3. $\alpha = 0.10,\ n = 12 - 1 = 11$
4. From table of Wilcoxon signed rank for two tail test,

   $\alpha = 0.10,\ n = 11,$ then $a = 13$
   We will reject H_0 if $\min(T^+, T^-) \leq a$

5. Since $\min(8.5, 57.5) = 8.5 \leq 14$, thus we reject H_0 and conclude that the population median for $d_i = x_i - y_i$ is not equal to zero.

Exercise:

A researcher conducts a pilot study to compare two treatments to help obese female teenagers lose weight. She tests each individual in two different treatment conditions. The data below provides the number of pounds that each participant lost.

Participant	Pounds Lost	
	Treatment 1	Treatment 2
1	10	18
2	20	12
3	15	16
4	9	7
5	18	21
6	11	17
7	6	13
8	12	14

Use the Wilcoxon signed ranks test at α=0.05 to determine that the two treatments are differ in losing weight.

Test de Mann-Whitney-Wilcoxon - Objectif

Objectif : comparaison de K = 2 échantillons indépendants par rapport a une variable X de nature :

- Quantitative
- Qualitative ordinale

Ce test regroupe 2 tests équivalents :Test U de Mann-Whitney et Test W de Wilcoxon

Mann-Whiteny Test

Test de Mann-Whitney-Wilcoxon - Hypothèses

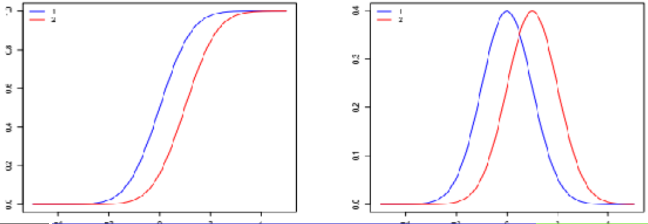
Soient $\begin{cases} F_1(X) \text{ la fonction de répartition de } X \text{ dans la population 1} \\ F_2(X) \text{ la fonction de répartition de } X \text{ dans la population 2} \end{cases}$

Les hypothèses de test sont :

$$\begin{cases} \mathcal{H}_0 \ : \ F_1(X) = F_2(X+\theta)\,;\ \theta = 0 & \text{Distributions identiques} \\ \mathcal{H}_1 \ : \ F_1(X) = F_2(X+\theta)\,;\ \theta \neq 0 & \text{Distributions différentes} \end{cases}$$

θ paramètre de translation : décalage entre les fonctions de répartition

Exemple de décalage $\theta \neq 0$



Mann-Whiteny Test

Test de Mann-Whitney-Wilcoxon - Statistique de test

Rappel : Somme de n premiers entiers

$$1 + 2 + \ldots + n = \frac{n(n+1)}{2}$$

Posons S_1 la somme des rangs des observations du groupe 1
Posons U_1 le nombre de couples $\{(x_{1i}, x_{2j}) \;/\; x_{1i} > x_{2j}\}$

$$U_1 = S_1 - \frac{n_1(n_1+1)}{2}$$

Posons S_2 la somme des rangs des observations du groupe 2
Posons U_2 le nombre de couples $\{(x_{1i}, x_{2j}) \;/\; x_{1i} < x_{2j}\}$

$$U_2 = S_2 - \frac{n_2(n_2+1)}{2}$$

Statistique de test : $U = \min(U_1, U_2)$

Mann-Whiteny Test

Test de Mann-Whitney-Wilcoxon - Statistique de test

Interprétation de la statistique de test $\rightarrow \mathcal{H}_0$ "totalement fausse" :

x x x x x x x o o o o o o o

rangs

$$\left\{\begin{array}{l} U_1 = \frac{n_1(n_1+1)}{2} - \frac{n_1(n_1+1)}{2} = 0 \\ U_2 = \sum_{i=n_1+1}^{n_1+n_2} r_i - \frac{n_2(n_2+1)}{2} = n_1 n_2 \end{array}\right\} U = U_1 = 0$$

o o o o o o o x x x x x x x

rangs

$$\left\{\begin{array}{l} U_1 = \sum_{i=n_2+1}^{n_1+n_2} r_i - \frac{n_1(n_1+1)}{2} = n_1 n_2 \\ U_2 = \frac{n_2(n_2+1)}{2} - \frac{n_2(n_2+1)}{2} = 0 \end{array}\right\} U = U_2 = 0$$

Mann-Whiteny Test

Test de Mann-Whitney-Wilcoxon - Statistique de test

Interprétation de la statistique de test $\rightarrow \mathcal{H}_0$ "Vraie" (mélange total) :

$S_p(n) =$ Somme des n premiers entiers pairs
$S_i(n) =$ Somme des n premiers entiers impairs

$$(1) \quad \xrightarrow{x\,o\,x\,o\,x\,o\,x\,o\,x\,o\,x\,o\,x\,o}_{rangs} \quad \begin{cases} U_1 = S_i(n_1) - \frac{n_1(n_1+1)}{2} \\ U_2 = S_p(n_2) - \frac{n_2(n_2+1)}{2} \end{cases}$$

$$(2) \quad \xrightarrow{o\,x\,o\,x\,o\,x\,o\,x\,o\,x\,o\,x\,o\,x}_{rangs} \quad \begin{cases} U_1 = S_p(n_1) - \frac{n_1(n_1+1)}{2} \\ U_2 = S_p(n_2) - \frac{n_2(n_2+1)}{2} \end{cases}$$

On peut en déduire que $\rightarrow$ Propriété : $U \leq \frac{n_1 n_2}{2}$

Sous $\mathcal{H}_0$, on montre que

$$\mathbb{E}[U] = \frac{n_1 n_2}{2} \quad \mathbb{V}[U] = n_1 n_2 \frac{n_1+n_2+1}{12}$$

Mann-Whiteny Test

- **Procédure.** On classe les variables $(X_1, ..., X_n; Y_1, ..., Y_m)$ par leur rang global et on note $R_1, R_2, ..., R_n \in \{1, 2, ..., n+m\}$ les rangs associés à l'échantillon X .
- **Exemple.** $X = (3, 5, 2)$; $Y = (1, 4)$ alors $Y_1 \leq X_3 \leq X_1 \leq Y_2 \leq X_2$ et les rangs associés à l'échantillon X sont $R_1 = 3; R_2 = 5; R_3 = 2.$
- **Idée.** Sous H_0, $R_1, ..., R_n$ sont uniformément répartis sur $\{1, 2, ..., n+m\}.$
- **Statistique.** Soit

$$\Sigma_1 = R_1 + ... + R_n \quad \text{et} \quad W_{YX} = \Sigma_1 - \frac{n(n+1)}{2}$$

- Propriété : W_{YX} est égal au nombre de paires (X_i, Y_j) (parmi les nm paires possibles) telles que $X_i =\geq Yj$.

Mann-Whiteny Test

thèoréme

Sous $H_0 : F = G$, la loi de W_{YX} est libre et symétrique par rapport à $\frac{nm}{2}$. De plus,

- $\mathbb{E}[W_{YX}] = \frac{nm}{2} \qquad V[W_{YX}] = \frac{nm(n+m+1)}{12}$
- $\frac{W_{YX} - \mathbb{E}[W_{YX}]}{\sqrt{V[W_{YX}]}} \rightsquigarrow \mathcal{N}(0,1)$
- *Test. On rejette H_0 pour les grandes valeurs de $[W_{YX} - \frac{nm}{2}]$.*
    - *Loi tabulée pour les petites valeurs de n et m (< 10).*
    - *Pour les grandes valeurs, on utilise l'approximation gaussienne.*

Mann-Whiteny Test

Example:

Data below show the marks obtained by electrical engineering students in an examination:

Gender	Marks
Male	60
Male	62
Male	78
Male	83
Female	40
Female	65
Female	70
Female	88
Female	92

Can we conclude the achievements of male and female students identical at significance level $\alpha = 0.1$

Mann-Whiteny Test

Solution:

Gender	Marks	Rank
Male	60	2
Male	62	3
Male	78	6
Male	83	7
Female	40	1
Female	65	4
Female	70	5
Female	88	8
Female	92	9

1. H_0 : Male and Female achievement are the same
   H_1 : Male and Female achievement are not the same

- We have $n_1 = 4$, $n_2 = 5$, $\Sigma = \sum R_i = 2 + 3 + 6 + 7 = 18$
- From the table of Wilcoxon rank sum test for $\alpha = 0.1$(two tail test) , $n_1 = 4$, $n_2 = 5$ so critical value [13,27]
- Reject H_0 if $w \neq [a, b]$ or $w \neq [13, 27]$

Mann-Whiteny Test

Exercise:

Using high school records, Johnson High school administrators selected a random sample of four high school students who attended Garfield Junior High and another random sample of five students who attended Mulbery Junior High. The ordinal class standings for the nine students are listed in the table below. Test using Mann-Whitney test at 0.05 level of significance.

Garfield J. High		Mulbery J. High	
Student	Class standing	Student	Class standing
Fields	8	Hart	70
Clark	52	Phipps	202
Jones	112	Kirwood	144
TIbbs	21	Abbott	175
		Guest	146

Kruskal-Wallis Test

ce test ;

- Une extension du test de Mann-Whiteny ou le test de la somme des rangs de Wilcoxon de la section précédente.
- Il compare plus de deux échantillons indépendants.
- C'est un test non paramétrique homologue de l'analyse de la variance(ANOVA).
- il ne suppose pas que l'échantillon a été tiré à partir de populations normalement distribuées avec des variances égales.

soit k echontillon indépendants, les hypothèse ont testé sont :

- $H_0 : M_1 = M_2 = ... = M_k$(the population median are equal) contre
- at least one M_i differs from the others the population median are not equal

Kruskal-Wallis Test

Test statistic H :

Kruskal–Wallis Test Based on *H* for Comparing *k* Population Distributions

Null hypothesis: H_0: The k population distributions are identical.

Alternative hypothesis: H_a: At least two of the population distributions differ in location.

Test statistic: $H = \{12/[n(n+1)]\}\sum_{i=1}^{k} R_i^2/n_i - 3(n+1)$, where

n_i = number of measurements in the sample from population i,
R_i = rank sum for sample i, where the rank of each measurement is computed according to its relative size in the overall set of $n = n_1 + n_2 + \cdots + n_k$ observations formed by combining the data from all k samples.

Rejection region: Reject H_0 if $H > \chi^2_\alpha$ with $(k-1)$ df.

Assumptions: The k samples are randomly and independently drawn. There are five or more measurements in each sample.

Kruskal-Wallis Test

Exemple :

In this case, $n_1 = 10 = n_2 = n_3$ and $n = 30$. Thus,

$$H = \frac{12}{30(31)}\left[\frac{(120)^2}{10} + \frac{(210.5)^2}{10} + \frac{(134.5)^2}{10}\right] - 3(31) = 6.097.$$

Table 15.6 Data for Example 15.7

Line 1		Line 2		Line 3	
Defects	Rank	Defects	Rank	Defects	Rank
6	5	34	25	13	9.5
38	27	28	19	35	26
3	2	42	30	19	15
17	13	13	9.5	4	3
11	8	40	29	29	20
30	21	31	22	0	1
15	11	9	7	7	6
16	12	32	23	33	24
25	17	39	28	18	14
5	4	27	18	24	16
	$R_1 = 120$		$R_2 = 210.5$		$R_3 = 134.5$

Kruskal-Wallis Test

Exemple :

Because all the n_i values are greater than or equal to 5, we may use the approximation for the null distribution of H and reject the null hypothesis of equal locations if $H > \chi^2_\alpha$ based on $k - 1 = 2$ df.
$\chi^2_{.05} = 5.99147$. Thus, we reject the null hypothesis at the $\alpha = .05$ level and conclude that at least one of the three lines tends to produce a greater number of defects than the others.

the value of $H = 6.097$ leads to rejection of the null hypothesis if $\alpha = .05$ but not if $\alpha = .025$. Thus, $.025 < p$-value $< .05$. The applet *Chi-Square Probability and Quantiles* can be used to establish that the approximate p-value $= P(\chi^2 > 6.097) = .0474$. ■

Kruskal-Wallis Test

Exemple :

Exercises

The table that follows contains data on the leaf length for plants of the same species at each of four swampy underdeveloped sites. At each site, six plants were randomly selected. For each plant, ten leaves were randomly selected, and the mean of the ten measurements (in centimeters) was recorded for each plant from each site. Use the Kruskal–Wallis H test to determine whether there is sufficient evidence to claim that the distribution of mean leaf lengths differ in location for at least two of the sites. Use $\alpha = .05$. Bound or find the approximate p-value.

Site	Mean Leaf Length (cm)					
1	5.7	6.3	6.1	6.0	5.8	6.2
2	6.2	5.3	5.7	6.0	5.2	5.5
3	5.4	5.0	6.0	5.6	4.0	5.2
4	3.7	3.2	3.9	4.0	3.5	3.6

Illustration sous R

Illustration sous R !

Exercises

Exercises

Exercises01 :Large-sample wilcoxon signed rank test

Test using level of significance $\alpha = 0.10$ whether the population median age of missing children differs from 6 years old, using the random sample of 50 missing children .10 children are 6 years old.

solution ;

1. H_0 : M= 6 versus $H_1 : M \neq 6$
2. Note that the original sample size (N) is 50 ,The ŞWilcoxon StatisticŤ is the value of $T_{data} = 279$, which represents the smaller of $T^+ = 279$ and $|T^-| = 541$
3. $Z_{data} = \dfrac{Z_{data} - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} = \dfrac{279 - \frac{40(40+1)}{4}}{\sqrt{\frac{40(41)(81)}{24}}} = -1.7608$
4. Because $Z_{data} \approx -1.7608 \leq -1.645$, we reject H0.

Exercises01 :Large-sample wilcoxon signed rank test

Test using level of significance $\alpha = 0.10$ whether the population median age of missing children differs from 6 years old, using the random sample of 50 missing children .10 children are 6 years old.

solution ;

1. H_0 : M= 6 versus $H_1 : M \neq 6$
2. Note that the original sample size (N) is 50 ,The ŞWilcoxon StatisticŤ is the value of $T_{data} = 279$, which represents the smaller of $T^+ = 279$ and $|T^-| = 541$
3. $Z_{data} = \dfrac{Z_{data} - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} = \dfrac{279 - \frac{40(40+1)}{4}}{\sqrt{\frac{40(41)(81)}{24}}} = -1.7608$
4. Because $Z_{data} \approx -1.7608 \leq -1.645$, we reject H0.

Exercises01 :Large-sample wilcoxon signed rank test

Test using level of significance $\alpha = 0.10$ whether the population median age of missing children differs from 6 years old, using the random sample of 50 missing children .10 children are 6 years old.

solution ;

1. H_0 : M= 6 versus $H_1 : M \neq 6$
2. Note that the original sample size (N) is 50 ,The ŞWilcoxon StatisticŤ is the value of $T_{data} = 279$, which represents the smaller of $T^+ = 279$ and $|T^-| = 541$
3. $Z_{data} = \dfrac{Z_{data} - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} = \dfrac{279 - \frac{40(40+1)}{4}}{\sqrt{\frac{40(41)(81)}{24}}} = -1.7608$
4. Because $Z_{data} \approx -1.7608 \leq -1.645$, we reject H0.

Exercises01 :Large-sample wilcoxon signed rank test

Test using level of significance $\alpha = 0.10$ whether the population median age of missing children differs from 6 years old, using the random sample of 50 missing children .10 children are 6 years old.

solution ;

1. H_0 : M= 6 versus $H_1 : M \neq 6$
2. Note that the original sample size (N) is 50 ,The ŞWilcoxon StatisticŤ is the value of $T_{data} = 279$, which represents the smaller of $T^+ = 279$ and $|T^-| = 541$
3. $Z_{data} = \dfrac{Z_{data} - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} = \dfrac{279 - \frac{40(40+1)}{4}}{\sqrt{\frac{40(41)(81)}{24}}} = -1.7608$
4. Because $Z_{data} \approx -1.7608 \leq -1.645$, we reject H0.

?

?