• <u>TEST ANOVA</u> :
Soient A un facteur à p modalités $A_1, \dots, A_p$ et X un caractère quantitatif, avec $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, on suppose que $\sigma_1^2 = \cdots = \sigma_p^2 = \sigma^2$.
On observe la valeur de X_i pour chaque n_i individus d'un échantillon indépendants.

$\mathcal{H}_0$: "$\forall i$; $\mu_i = \mu$" VS $\mathcal{H}_1$: "$\exists i$; $\mu_i \neq \mu$"

	sce (SS)	*ddl* (DF)	*cm* (MS)	f_{obs} (F)
Total	$\sum_{i=1}^{p}\sum_{j=1}^{n_i} x_{i,j}^2 - n\bar{x}^2 = sce_F + sce_R$	$n-1$		
Factoriel	$\sum_{i=1}^{p} n_i \bar{x_i}^2 - n\bar{x}^2$	$p-1$	$\frac{sce_F}{ddl_F}$	$\frac{cm_F}{cm_R}$
Résiduel	$\sum_{i=1}^{p}\sum_{j=1}^{n_i} x_{i,j}^2 - \sum_{i=1}^{p} n_i \bar{x_i}^2$	$n-p$	$\frac{sce_R}{ddl_R}$	

On calcule f_α tq $\mathbb{P}(F \geq f_\alpha) = \alpha$, où $F \sim \mathcal{F}(p-1, n-p)$. On a $\mathcal{R}_\alpha = [f_\alpha ; \infty[$.

Ou si $p-\text{value} = \mathbb{P}(F \geq f_{obs}) \leq \alpha$, alors on rejette $\mathcal{H}_0$.

• <u>TEST ANOVA (égalité des variances)</u> : $s_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}\left(x_{i,j} - \bar{x_i}\right)^2 = \frac{1}{n_i-1}\left(\sum_{j=1}^{n_i} x_{i,j}^2 - n_i \bar{x_i}^2\right)$ et $s^2 = \frac{1}{n-p}\sum_{i=1}^{p}(n_i - 1)s_i^2$

$\mathcal{H}_0$: "$\forall i$; ${\sigma_i}^2 = \sigma^2$" VS $\mathcal{H}_1$: "$\exists i$; ${\sigma_i}^2 \neq \sigma^2$"

Tests	Statistique de test	Quantile	Zone rejet $\mathcal{R}_\alpha$
Bartlett (si $\min(n_1, \dots, n_p) \geq 4$)	$K = \dfrac{(n-p)\ln(s^2) - \sum_{i=1}^{p}(n_i - 1)\ln(s_i^2)}{1 + \frac{1}{3(p-1)}\left(\sum_{i=1}^{p}\frac{1}{n_i - 1} - \frac{1}{n-p}\right)}$	k_α tq $\mathbb{P}(K \geq k_\alpha) = \alpha$, où $K \sim \chi^2(p-1)$	$[k_\alpha ; +\infty[$
Cochran (si $n_1 = \cdots = n_p$)	$C_{obs} = \frac{max(s_i^2)}{\sum_{i=1}^{p} s_i^2}$	$c(m,p)$ $(voir\ table)$	$[c(m,p) ; +\infty[$

• <u>TEST ANOVA (comparaison 2 moyennes μ_k et μ_ℓ)</u> : $s_R = \sqrt{cm_R}$

$\mathcal{H}_0$: "$\mu_k = \mu_\ell$" VS $\mathcal{H}_1$: "$\mu_k \neq \mu_\ell$"

Test	Statistique de test	Quantile	Zone rejet $\mathcal{R}_\alpha$
Bonferroni	$T = \dfrac{\overline{x_k} - \overline{x_\ell}}{s_R\sqrt{\frac{1}{n_k} + \frac{1}{n_\ell}}}$	t_α^{**} tq $\mathbb{P}(\lvert T\rvert \geq t_\alpha^{**}) = \frac{2\alpha}{p(p-1)}$, où $T \sim \mathcal{S}tu(n-p)$	$]-\infty ; -t_\alpha^{**}] \cup [t_\alpha^{**} ; +\infty[$

LOI GRANDS NOMBRES : Si on répète N fois une expérience avec une proba p d'apparition d'un événement A, la fréquence de cet événement tend vers p quand $N \longrightarrow \infty$.

Si p-val ≥ α ⇒ non-rejet $\mathcal{H}_0$
Si p-val < α ⇒ rejet $\mathcal{H}_0$

Cas unilatéral gauche : $p-val = \mathbb{P}(X \leq x_{obs})$
Cas unilatéral droit : $p-val = 1 - \mathbb{P}(X \leq x_{obs})$
Cas bilatéral : $p-val = 2\left(1 - \mathbb{P}(X \leq x_{obs})\right)$

	$\mathcal{U}(1/n)$	$\mathcal{B}er(p)$	$\mathcal{B}(n,p)$	$\mathcal{P}(\lambda)$	$\mathcal{U}([a,b])$	$\chi^2(n)$	$\mathcal{S}tu(n)$
$\mathbb{E}$	$(n+1)/2$	p	np	λ	$(b+a)/2$	n	$\begin{cases} 0, \text{si } n > 1 \\ \text{FI sinon} \end{cases}$
$\mathbb{V}$	$(n^2-1)/12$	$p(1-p)$	$np(1-p)$	λ	$(b-a)^2/12$	$2n$	$\begin{cases} \frac{n}{n-2}, \text{si } n > 2 \\ +\infty, \text{si } 1 < n \leq 2 \\ \text{FI sinon} \end{cases}$

Soit $X \sim \mathcal{B}(n,p)$. Si $n \geq 30$ et $np \leq 10$, alors $X \sim \mathcal{P}(\lambda)$, avec $\lambda = np$.
Soit $X \sim \mathcal{B}(n,p)$. Si $np \geq 10$ et $n(1-p) \geq 10$, alors $X \sim \mathcal{N}\left(np, \sqrt{np(1-p)}\right)$.
Soit $X \sim \mathcal{P}(\lambda)$. Si $\lambda \geq 10$, alors $X \sim \mathcal{N}\left(\lambda, \sqrt{\lambda}\right)$.

Variable qualitative → nominale = plusieurs modalités non mesurables s'excluant mutuellement / ordinale = degrés d'un état
Variable quantitative → résultat d'une mesure ou d'un comptage

---

$Y = \hat{a}X + \hat{b} \leftrightarrow \hat{a} = \frac{\mathrm{cov}(x,y)}{s_{\mathrm{ech}}^2(x)} = r\frac{s_{\mathrm{ech}}(y)}{s_{\mathrm{ech}}(x)}$ et $\hat{b} = \bar{y} - \hat{a}\bar{x}$ / $X = \hat{a}Y + \hat{b} \leftrightarrow \hat{a} = \frac{\mathrm{cov}(x,y)}{s_{\mathrm{ech}}^2(y)} = r\frac{s_{\mathrm{ech}}(x)}{s_{\mathrm{ech}}(y)}$ et $\hat{b} = \bar{y} - \hat{a}\bar{x}$ ⇒ droites MCO
angle + ouvert entre 2 MCOS ⇒ liaison moins fortes
Si nouvelle valeur x^* de X, on prédit $\widehat{y^*} = \hat{a}x^* + \hat{b}$

Etude des résidus

On appelle valeur ajustée de la $i^{\text{ème}}$ observation de la variable Y l'approximation

$$\hat{y}_i = \hat{a}x_i + \hat{b}$$

On appelle résidu e_i, l'erreur observée que l'on commet en approchant y_i par $\hat{y}_i$: $e_i = y_i - \hat{y}_i$

<u>Démarche</u> Prévision avec la droite des MCO

- calcul de la droite des MCO
- validation du modèle ⇒ étude des résidus et détection des valeurs aberrantes et influentes
- qualité de l'ajustement ⇒ décomposition de la variance, coefficient de détermination et test significativité globale
- qualité de prédiction (PRESS) et prédiction

Validité du modèle

- vérifier la normalité des résidus
- vérifier que les résidus ne contiennent pas d'information structurée
- vérifier que les résidus ne sont pas auto-corrélés entre eux

<u>Equation d'analyse de la variance</u>

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$$

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{\substack{\text{Somme des carrés}\\ \text{totale}\\ \text{SCT}}} = \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}_{\substack{\text{Somme des carrés}\\ \text{expliquée}\\ \text{SCE}}} + \underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{\substack{\text{Somme des carrés}\\ \text{résiduelle}\\ \text{SCR}}}$$

$$\underbrace{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}_{\substack{\text{Variance}\\ \text{totale}}} = \underbrace{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}_{\substack{\text{Variance}\\ \text{expliquée}}} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{\substack{\text{Variance}\\ \text{résiduelle}}}$$

Droite de Henry :

$P(\epsilon \leq e_i)$	0.077	0.154	0.231	0.308	0.385	0.462	0.538
e_i	-5.75	-4.27	-2.76	-2.75	-0.29	-0.25	1.19
u_i	-1.43	-1.02	-0.74	-0.50	-0.29	-0.10	0.10

<u>Ex :</u> $P(Z \leq u_i) = P(\epsilon \leq e_i) = \frac{7}{13} = 0.538 \Rightarrow u_i = 0.0954$

Coefficient de détermination

Part de la variance de y expliquée par la relation $\hat{y} = \hat{a}x + \hat{b}$

$$R^2 = \frac{\mathrm{Var}(\hat{y})}{\mathrm{Var}(y)}$$

Dans le cas d'un ajustement linéaire, on peut montrer que $R^2 = r^2(x,y)$ (où r est le coefficient de corrélation linéaire)

- $R^2 \in [0,1]$
- Plus R est proche de 1, plus le modèle explique correctement la variabilité de Y.

Croisement de deux variables quantitatives

Représentation graphique (nuage de points)

Coefficient de corrélation

- Calcul de l'indicateur statistique
- Test de nullité du coefficient de corrélation

Régression linéaire

- Estimation des coefficients
- Validité du modèle (Etude des résidus et des observations influentes)
- Qualité d'ajustement (R^2, significativité globale)
- Prédiction

<u>Propriété</u> : Un estimateur sans biais de la variance de l'erreur du modèle est

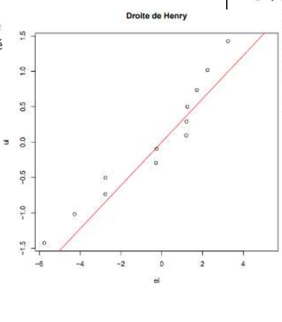
$$s_e^2 = \frac{1}{n-2}\sum e_i^2$$

Validité du modèle

- vérifier la normalité des résidus
- vérifier que les résidus ne contiennent pas d'information structurée
- vérifier que les résidus ne sont pas auto-corrélés entre eux

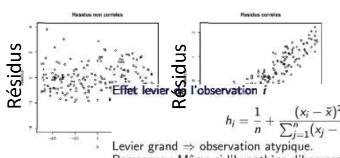
<u>Vérification de la normalité des résidus</u>

- histogramme ⇒ la distribution doit être unimodale et symétrique autour de 0.
- tests (Kolmogorov-Smirnov, Shapiro Wilks, ...)
- droite de Henry ⇒ confronte les quantiles théoriques de normale et la distribution cumulée estimée sur les donné



<u>Vérification de l'homoscédasticité des résidus</u>

Les résidus sont homoscédastiques si leur répartition est homogène et ne dépend pas des valeurs de la variable explicative (et donc pas non plus des valeurs prédites).
On vérifie que les résidus n'ont pas de structure particulière en traçant un graphe des résidus :



Effet levier de l'observation i

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}$$

Levier grand ⇒ observation atypique.
<u>Remarque</u> : Même si l'hypothèse d'homoscédasticité est vérifiée, les résidus n'ont pas la même variance.
$E(e_i) = 0$ et $Var(e_i) = \sigma^2(1 - h_i)$

Tableau de distribution des fréquences

Modalités	Effectif	Fréquence
x_1	n_1	$f_1 = n_1/n$
$\vdots$	$\vdots$	$\vdots$
x_p	n_p	$f_p = n_p/n$
Total	n	1

On définit l'angle α_i de la modalité x_i par $\alpha_i = 360 \times f_i$.

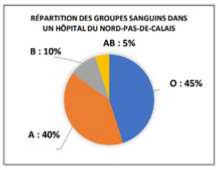


Tableau de contingence

	T	Somme
P	$n_{11} \cdots n_{1j} \cdots n_{1t}$ $\vdots \quad \vdots \quad \vdots$ $n_{i1} \cdots n_{ij} \cdots n_{it}$ $\vdots \quad \vdots \quad \vdots$ $n_{p1} \cdots n_{pj} \cdots n_{pt}$	$n_{1\star}$ $\vdots$ $\vdots$ $\vdots$ $n_{p\star}$
Somme	$n_{\star 1} \cdots \cdots \cdots n_{\star t}$	n

Variables qualitative nominale

Biais : $B\left(\widehat{\theta_n}\right) = \mathbb{E}\left(\widehat{\theta_n} - \theta\right)$

Tableau de distribution des effectifs et fréquences cumulés

Valeur des modalité	Effectif n_i	Fréquence f_i	Effectif cumulé N_i	Fréquence cumulée F_i	Fréquence cumulée G_i
x_1	n_1	f_1	n_1	f_1	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	n_i	f_i	$n_1 + \cdots + n_i$	$f_1 + \cdots + f_i$	$f_p + \cdots + f_{p-i}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_p	n_p	f_p	n	1	f_p

Variables qualitative ordinale et quantitative discrète

Tableau de distribution des effectifs et fréquences cumulés

Classes	Effectif n_i	Fréquence f_i	Densité	Effectif cumulé N_i	Fréquence cumulée F_i	Fréquence cumulée G_i
$]a_1, b_1]$	n_1	f_1	$d_1 = f_1/(b_1 - a_1)$	n_1	f_1	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$]a_i, b_i]$	n_i	f_i	$d_i = f_i/(b_i - a_i)$	$n_1 + \cdots + n_i$	$f_1 + \cdots + f_i$	$f_p + \cdots + f_{p-i}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$]a_p, b_p]$	n_p	f_p	$d_p = f_p/(b_p - a_p)$	n	1	f_p

Variables quantitatives continue

Intervalle confiance moyenne si σ inconnue : $\left[\bar{x} - t_{1-\alpha/2}^{n-1}\frac{\hat{\sigma}}{\sqrt{n}} ; \bar{x} + t_{1-\alpha/2}^{n-1}\frac{\hat{\sigma}}{\sqrt{n}}\right]$ / si σ connue : $\left[\bar{x} - u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} ; \bar{x} + u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$

Conditions : Si $n \leq 30$: il faut que $X \sim \mathcal{N}$ / Si $n > 30$: pas de condition sur la loi de X.

Intervalle confiance proportion : on pose $\hat{\pi} = p$, si $n\hat{\pi} \geq 10$ et $n(1 - \hat{\pi}) \geq 10$: $\left[\hat{\pi} - u_{1-\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{\mathrm{n}}} ; \hat{\pi} + u_{1-\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{\mathrm{n}}}\right]$

2 variables qualitatives ⇒ test χ^2 | 2 variables quantitatives ⇒ test coeff corrélation | 1 quali & 1 quanti ⇒ t-test ou ANOVA

Test comparaison proportion à π_0 connue (1 échantillon) : $U = \sqrt{n}\frac{p - \pi_0}{\sqrt{\pi_0(1-\pi_0)}} \sim \mathcal{N}(0,1)$ / $n_1, n_2 \geq 30$ et $n\pi_0 \geq 5$ et $n(1 - \pi_0) \geq 5$.

Test comparaison proportions (2 échantillons) : $U = \frac{p_1 - p_2}{\sqrt{\frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2}}} \sim \mathcal{N}(0,1)$, où $p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$ / $n_1, n_2 \geq 30$ et $\min\limits_{i=1,2}(n_i p ; n_i(1-p)) \geq 5$

Test comparaison moyenne à μ_0 connue (1 échantillon) : $T = \sqrt{n}\frac{\bar{x} - \mu_0}{\hat{\sigma}} \sim \mathcal{S}tu(n-1)$ / $n \geq 30$ ou $X \sim \mathcal{N}(\mu, \sigma)$

Test Fisher (comparaison 2 variances observées s_1^2, s_2^2) : $F = \frac{\widehat{\sigma_1}^2}{\widehat{\sigma_2}^2} \sim \mathcal{F}(n_1 - 1 ; n_2 - 1)$ / $n_1, n_2 \geq 30$ ou $X_{1,2} \sim \mathcal{N}(\mu_{1,2}, \sigma_{1,2})$

t-test comparaison moyennes (2 échantillons indépendants) : $T = \frac{\bar{x}_1 - \bar{x}_2}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim \mathcal{S}tu(n_1 + n_2 - 2)$, où $S = \sqrt{\frac{(n_1 - 1)\widehat{\sigma_1}^2 + (n_2 - 1)\widehat{\sigma_2}^2}{n_1 + n_2 - 2}}$ si $\sigma_1^2 = \sigma_2^2$ / $n_1, n_2 \geq 30$ ou $X_{1,2} \sim \mathcal{N}(\mu_{1,2}, \sigma_{1,2})$

$T = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\widehat{\sigma_1}^2}{n_1} + \frac{\widehat{\sigma_2}^2}{n_2}}} \sim \mathcal{S}tu(\nu)$, si $\sigma_1^2 \neq \sigma_2^2$ (Si NON conditions ⇒ test Mann & Whitney)

t-test séries appariées :

Sujets	1	2	$\cdots$	n
X_A	X_1^A	X_2^A	$\cdots$	X_n^A
X_B	X_1^B	X_2^B	$\cdots$	X_n^B
$D = X_A - X_B$	$D_1 = X_1^A - X_1^B$	$D_2 = X_2^A - X_2^B$	$\cdots$	$D_n = X_n^A - X_n^B$

$T = \sqrt{n}\frac{\bar{d}}{\widehat{\sigma_D}^2} \sim \mathcal{S}\mathrm{tu}(\mathrm{n} - 1)$ / $n \geq 30$ ou $D \sim \mathcal{N}(\mu_D, \sigma_D)$

(Si NON conditions ⇒ Test Wilcoxon)

Test χ^2 ajustement (1 quali / 1 échantillon) | Test χ^2 indépendance (2 qualis / 1 échantillon) | Test χ^2 homogénéité (1 quali / l >2 échantillons)

Modalités	1	2	...	c
O_i	O_1	O_2	...	O_c
π_i	π_1	π_2	...	π_c
T_i	$n \times \pi_1$	$n \times \pi_2$	...	$n \times \pi_c$

$K = \sum_{i=1}^{c}\frac{O_i^2}{T_i} - n \sim \chi^2(c-1)$ / $\forall i$; $T_i \geq 5$

(Si NON conditions ⇒ regrouper des T_i si possible)

		Y				Somme
		1	2	$\cdots$	c	
X	1	O_{11}	O_{12}	$\cdots$	O_{1c}	n_1
	2	O_{21}	O_{22}	$\cdots$	O_{2c}	n_2
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	l	O_{l1}	O_{l2}	$\cdots$	O_{lc}	n_l
Somme		m_1	m_2	$\cdots$	m_c	n

		X				Somme
		1	2	$\cdots$	c	
Echantillon	1	O_{11}	O_{12}	$\cdots$	O_{1c}	n_1
	2	O_{21}	O_{22}	$\cdots$	O_{2c}	n_2
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	l	O_{l1}	O_{l2}	$\cdots$	O_{lc}	n_l
Somme		m_1	m_2	$\cdots$	m_c	n

Test de Mann & Whitney :

- Etape 0 : $\mathcal{H}_0$: "$M_{ech1} = M_{ech2}$" VS $\mathcal{H}_1$: "$M_{ech1} \neq M_{ech2}$"
- Etape 1 : Ranger par ordre croissant l'ensemble des 2 échantillons
- Etape 1bis : Déterminer les rangs. Si doublon, faire rang moyen
- Etape 2 : Calcul de $T_1 = \sum \mathrm{rang}_{ech1}$ et $T_2 = \sum \mathrm{rang}_{ech2}$
- Etape 3 : Calcul de $U_{12} = \mathrm{T}_2 - \frac{\mathrm{n}_2(\mathrm{n}_2+1)}{2}$ et $U_{21} = T_1 - \frac{n_1(n_1+1)}{2}$
- Etape 4 : Statistique de test $U = \min(U_{12} ; U_{21})$
- Etape 5 : Recherche de la zone de rejet sur la table

Si $n_1, n_2 \geq 20$, on pose $Z = \frac{U - \mu}{\sigma} \sim \mathcal{N}(0,1)$ avec $\mu = \frac{n_1 n_2}{2}$ et $\sigma = \sqrt{\frac{n_1 n_2(n_1 + n_2 + 1)}{12}}$

Réécriture de la zone de rejet avec Z.

Test de Wilcoxon

- Tableau

Sujets	1	2	$\cdots$	n
X_A	X_1^A	X_2^A	$\cdots$	X_n^A
X_B	X_1^B	X_2^B	$\cdots$	X_n^B
$D = X_A - X_B$	$D_1 = X_1^A - X_1^B$	$D_2 = X_2^A - X_2^B$	$\cdots$	$D_n = X_n^A - X_n^B$

- Etape 0 : $\mathcal{H}_0$: "$M_D = 0$" VS $\mathcal{H}_1$: "$M_D \neq 0$" (ou "$M_D > 0$" ou "$M_D < 0$")
- Etape 1 : Eliminer les $D_i = 0$
- Etape 2 : Ranger par ordre croissant l'ensemble des $|D_i|$
- Etape 2bis : Déterminer les rangs. Si doublon, faire rang moyen.
- Etape 3 : Calcul de $P = \sum \mathrm{rang}_{D_i > 0}$ et $M = \sum \mathrm{rang}_{D_i < 0}$
- Etape 4 : Statistique de test : $T = \min(M ; P)$
- Etape 5 : Zone de rejet $\mathcal{R}_\alpha = \{T \leq t\}$

Si $n_1, n_2 \geq 20$, on pose $Z = \frac{T - \mu}{\sigma} \sim \mathcal{N}(0,1)$ avec $\mu = \frac{n(n+1)}{4}$ et $\sigma = \sqrt{\frac{n(n+1)(2n+1)}{24}}$

Réécriture de la zone de rejet avec Z.

Test coefficient corrélation : $T = \frac{r(X,Y)\sqrt{n-2}}{\sqrt{1 - r(X,Y)^2}} \sim \mathcal{S}tu(n-2)$ / $X, Y \sim \mathcal{N}(\mu_{X,Y}, \sigma_{X,Y})$ / $\mathcal{R}_\alpha = \left]-\infty ; -t_{1-\alpha/2}^{n-2\,\mathrm{ddl}}\right] \cup \left[t_{1-\alpha/2}^{n-2\,\mathrm{ddl}} ; \infty\right[$

Test