P-value function

Paris School of Business

11/14/2020

Le package pvaluefunctions:

Le package pvaluefunctions permet de créer des graphiques des fonctions de la p-value , des distributions de confiance, des densités de confiance ou de la valeur Surprisale (valeur S) (Groenland 2019). Installation Vous pouvez installer le package directement en tapant install.packages("pvaluefunctions"). Après l'installation, chargez-le dans R en utilisant library(pvaluefunctions). La fonction dépend des packages R suivants, qui doivent être installés au préalable:

- ggplot2
- Balance
- zipfR

Utilisez la commande install.packages(c("ggplot2", "scales", "zipfR"))dans R pour installer ces packages. les données que nous allons utiliser sont issues de la table célèbre, collectées par Edgar Anderson. le fichier donne les mesures en centimètre des variables suivantes:

- longueur du sépale (Sepal.Length)
- largeur du sépale (Sepal.Width)
- longueur du pétale (Petal.Length)
- largeur du pétale (Petal.Width)

pour trois espèces d'iris qui sont les :

- Iris setosa
- Iris versicolor
- Iris virginica

Data IRIS:

pour importer cette table de données, il suffit d'importer la librairie "datasets" et puis exécuter l'instruction "data(iris)". T-test Pour tester l'égalité de la moyenne des sepal.length de l'espèce setosa avec celle de l'espèce versicolor, on fait appel au t-test, à deux échantillons avec deux variances inégales, aussi connu par le welch test. pour cela, il est indispensable que les données sur lesquelles nous allons appliquer le test respect 3 règles: -les échantillons sont normalement distribués, -l'écart type des deux populations est inconnu et suppose qu'il est inégal, l'échantillon est suffisamment grand (sur 30). Notre objectif à travers cette étude, est d'expliquer l'emploi et l'utilité du package pvaluefunctions dans la construction des graphiques des fonctions p-value, distributions de confiance et les densités de confiance

```
#-------------------------------------------------------------------------
#importation du package "pvaluefunctions".
#-------------------------------------------------------------------------

library(pvaluefunctions)
```

```
## Warning: package 'pvaluefunctions' was built under R version 3.6.2
```

```r
#-------------------------------------------------------------------------------
#importation de la librarie "datasets"
#Base de données iris
#-------------------------------------------------------------------------------

library(datasets)
data(iris)
View(iris)
#On supprime l'espèce de type "virginica" de notre table.
iris <- iris[iris$Species != "virginica",]
iris$Species<- factor(iris$Species)

#calcul de la difference entre la moyenne de longueur du sépale type setosa et versicolor.
with(iris, mean(Petal.Length[ Species== "setosa"])) - with(iris, mean(Petal.Length[Species ==   "versicolor" ]))
```

```
## [1] -2.798
```

```r
#-------------------------------------------------------------------------------
# Application du test welch
#-------------------------------------------------------------------------------

# Hypothese nulle H0: mu(setosa) = (versicolor) vs hypothese alternative H1 mu(setosa) != mu(versicolor)
Test_welch<-t.test(Petal.Length ~ Species, data = iris, paired=TRUE,var.equal = FALSE)
Test_welch
```

```
##
##  Paired t-test
##
## data:  Petal.Length by Species
## t = -37.241, df = 49, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.948983 -2.647017
## sample estimates:
## mean of the differences
##                  -2.798
```

Interprétation des Résultats du Test du welch

Comme nous pouvons l'observer d'après les résultats ci-dessus, la p-value qui est de 2.2e-16 est nettement inférieure à notre seuil de 5%. Par conséquence, nous pouvons rejeter l'hypothèse nulle et accepter l'alternative, en d'autres termes la longueur du sépale à un certain effet sur les espèces.

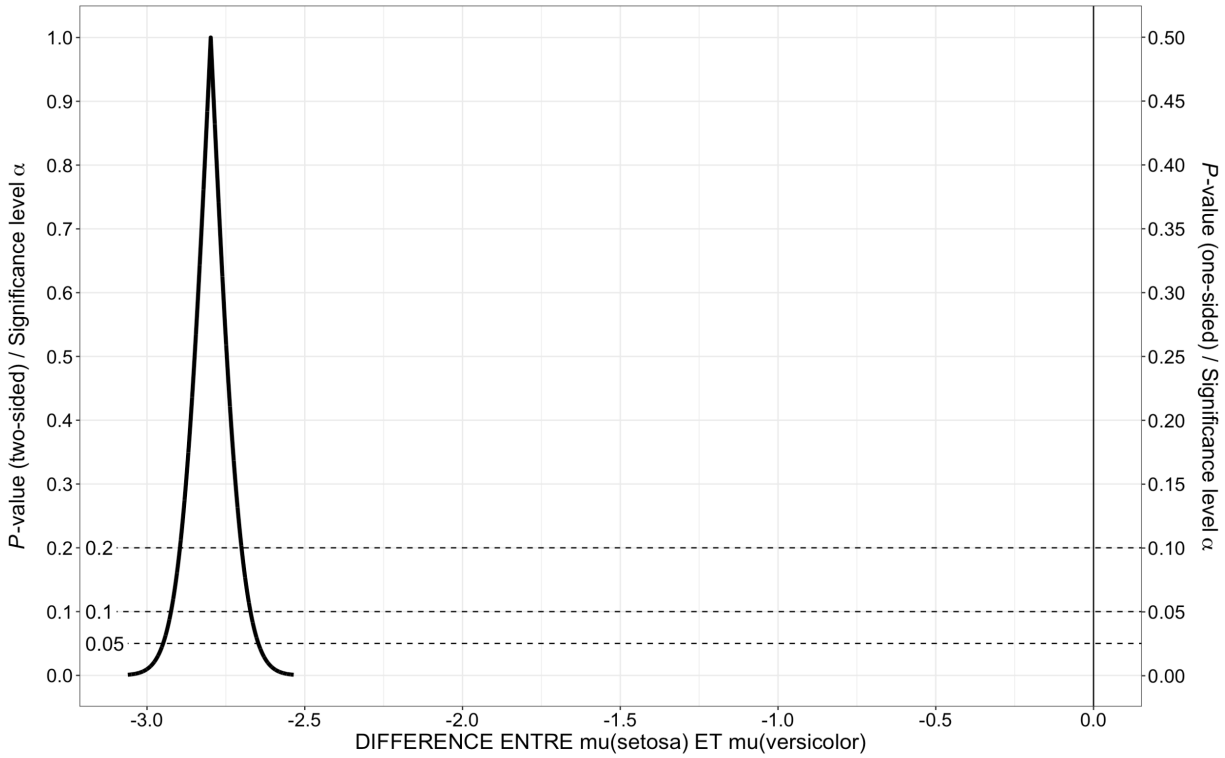
Graphe

Maintenant, nous allons illustrer la pvalue par le graphe suivant ,en tenant compte de trois paramètres qui sont:

- estimate qui indique de notre cas, la différence entre les deux moyennes estimées
- df qui indique le degré de liberté
- tstat est la statistique de décision du test de welch

```r
# Le graphe (conf_dist)

res <- conf_dist(
  estimate = c(-2.798)
  , df = c(49)
  , tstat = c(-37.241)
  , type = "ttest"
  , plot_type = "p_val"
  , n_values = 1e4L
```

```
  # , est_names = c("")
  , conf_level = c(0.95, 0.90, 0.80)
  , null_values = c(0)
  , trans = "identity"
  , alternative = "two_sided"
  , log_yaxis = FALSE
  , cut_logyaxis = 0.05
  , xlab = "DIFFERENCE ENTRE mu(setosa) ET mu(versicolor)"
  , together = FALSE
  , plot_p_limit = 1 - 0.999
  , plot_counternull = TRUE
  , x_scale = "line"
  , plot = TRUE)
```



Test du chi-carré

Vérifiant maintenant ,s'il y a une interaction entre Species et Petal.Width.Cat. Tout d'abord, nous allons convertir la variable continue "Petal.Width" en une variable catégorielle qui se nommera "Petal.Width.Cat", une fois ces deux variables sont catégorielles, nous utiliserons le test du chi-carré et ensuite nous élaborons un tableau de contingence qui va nous servir en moment du test.

NB : pour faire le test de chi-carré les variables doivent êtres qualitatives...

pour effectuer le test, nous supposerons les hypothèses ci-dessous:

-H0: Le Petal.Width.Cat n'a aucun effet sur l'espèce. vs H1: Le Petal.Width.Cat a un certain effet sur les espèces

```
# la function summary permet d'avoir la description statistique de la variable Petal.Width(largeur du pétale)
summary(iris$Petal.Width)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.100   0.200   0.800   0.786   1.300   1.800
```

```r
#convertion de la variable continu Petal.Width en variable catégorielle
iris$Petal.Width.Cat <- cut(iris$Petal.Width, breaks = quantile(iris$Petal.Width, probs = seq(0, 1, 0.5)),
                            include.lowest = TRUE)
levels(iris$Petal.Width.Cat) <- c("Dessous", "Dessus")
# "dessous" est affecté aux valeurs qui sont au dessous de la médiane et "dessus" l'inverse.
quantile(iris$Petal.Width, probs = seq(0, 1, 0.5))
```

```
##   0%  50% 100%
##  0.1  0.8  1.8
```

```r
#-------------------------------------------------------------------------
#création d'un tableau de contingence
#-------------------------------------------------------------------------

table_contingence <- table(iris$Petal.Width.Cat, iris$Species)
#application du test de KHI-DEUX.
Xsqt <- chisq.test(table_contingence)
Xsqt
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table_contingence
## X-squared = 96.04, df = 1, p-value < 2.2e-16
```

Interprétation des Résultats du Test du chi-carré

D'après les résultats, la p-value est inferieure à 2.2e-16 , ce qui est bien plus petit que la valeur du seuil de 5%. Cela nous permet de rejeter en l'hypothèse nulle et d'accepter l'hypothèse alternative. En d'autres termes, Petal.Width.Cat à un impact sur les espèces, ce qui nous permet de conclure que Petal.Width.Cat est un bon prédicteur pour les espèces.

La regression

Dans l'exemple qui suit nous tentons un premier modèle de régression multiple qui nous permettra de voir la relation entre le Sepal.Length comme variable à expliquer (output) et les prédicteurs suivants (variables explicatives): -Sepal.Width -Petal.Length -Petal.Width Pour cela nous allons utiliser la fonction lm et fonction summary pour afficher les résultats de notre modèle.

```r
modele_regression <- lm(Sepal.Length~Sepal.Width + Petal.Length+Petal.Width, data = iris)
summary(modele_regression )
```

```
##
## Call:
## lm(formula = Sepal.Length ~ Sepal.Width + Petal.Length + Petal.Width,
##     data = iris)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.76316 -0.23722  0.00556  0.18770  0.59723
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.07067    0.31620   6.549 2.86e-09 ***
## Sepal.Width   0.61196    0.07879   7.767 8.80e-12 ***
## Petal.Length  0.63897    0.10560   6.051 2.77e-08 ***
## Petal.Width  -0.41249    0.26325  -1.567     0.12
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2973 on 96 degrees of freedom
```

```
## Multiple R-squared:  0.7918, Adjusted R-squared:  0.7853
## F-statistic: 121.7 on 3 and 96 DF,  p-value: < 2.2e-16
```

Interprétation des Résultats de la regression

D'après la sortie, on constate que notre modèle à un pouvoir explicative de 78,53% (R ajusté = 78,53%),donc nous pouvons dire que notre modèle à priori est adapté pour prédire la variable Sepal.Length, en d'autres termes, la relation linéaire positive entres les variables (variable explicatives et variables à expliquer sepal.length) est assez forte (proche de 1)

Avant d'interpréter les résultats, il convient d'évaluer la significativité globale du modèle.

-Hypothèse du test de significativité globale du modèle

H0 : Absence de significativité globale des variables, H1: Au moins une variable n'est pas significativement différente de zéro.

Ici, comme la p-value associée au modèle(p-value=2.2e-16) est inférieure à 5%, on rejette H0, et on peut conclure que le modèle est globalement significatif.Si cela n'avait pas été le cas, il convient d'exclure à chaque fois les variables statistiquement non significatives (p-value associée à la variable>5%) , et refaire à chaque fois le test, jusqu'à ce que le modèle soit globalement significatif.

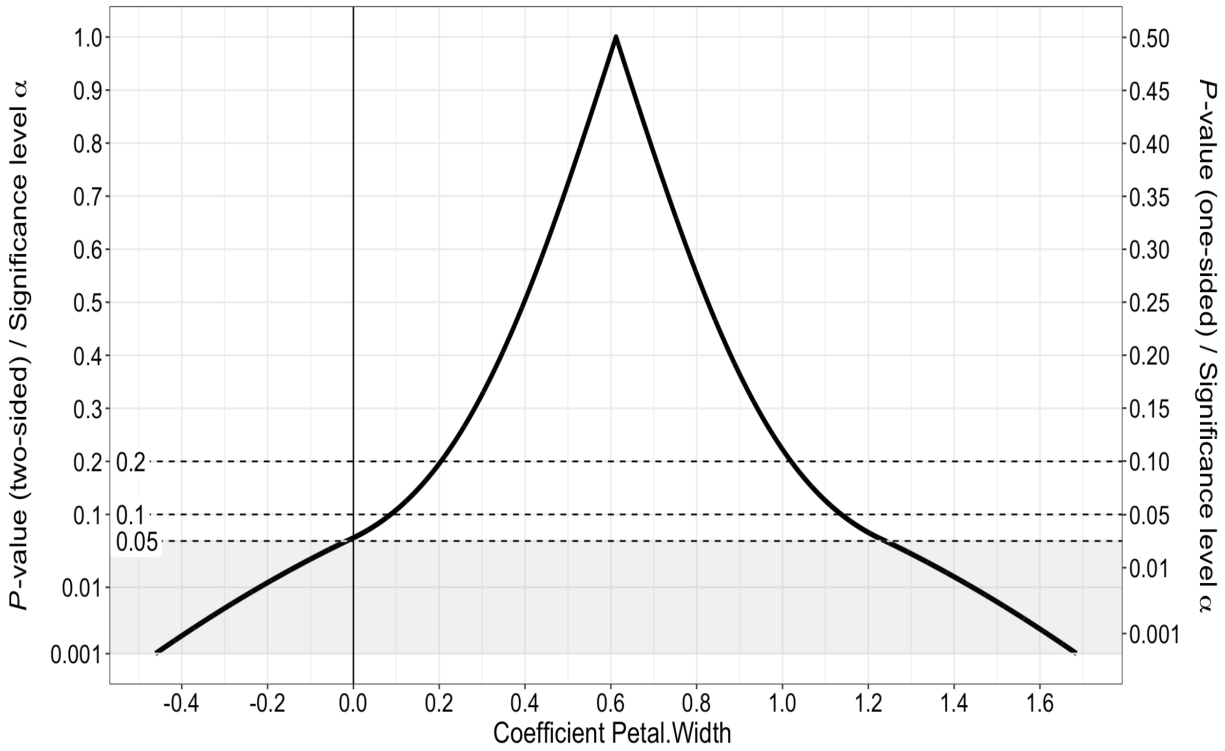
-Significativité des coefficients

D'après les résultats de la régression multiple , la p-value associé a la variable Sepal. with est inférieure au seuil 5%, donc nous pouvons conclure que le coefficient associé est significativement diffèrent de zéro.

les graphes ci-dessous nous montrent ces derniers résultats:

Graphe SEPAL.WIDTH

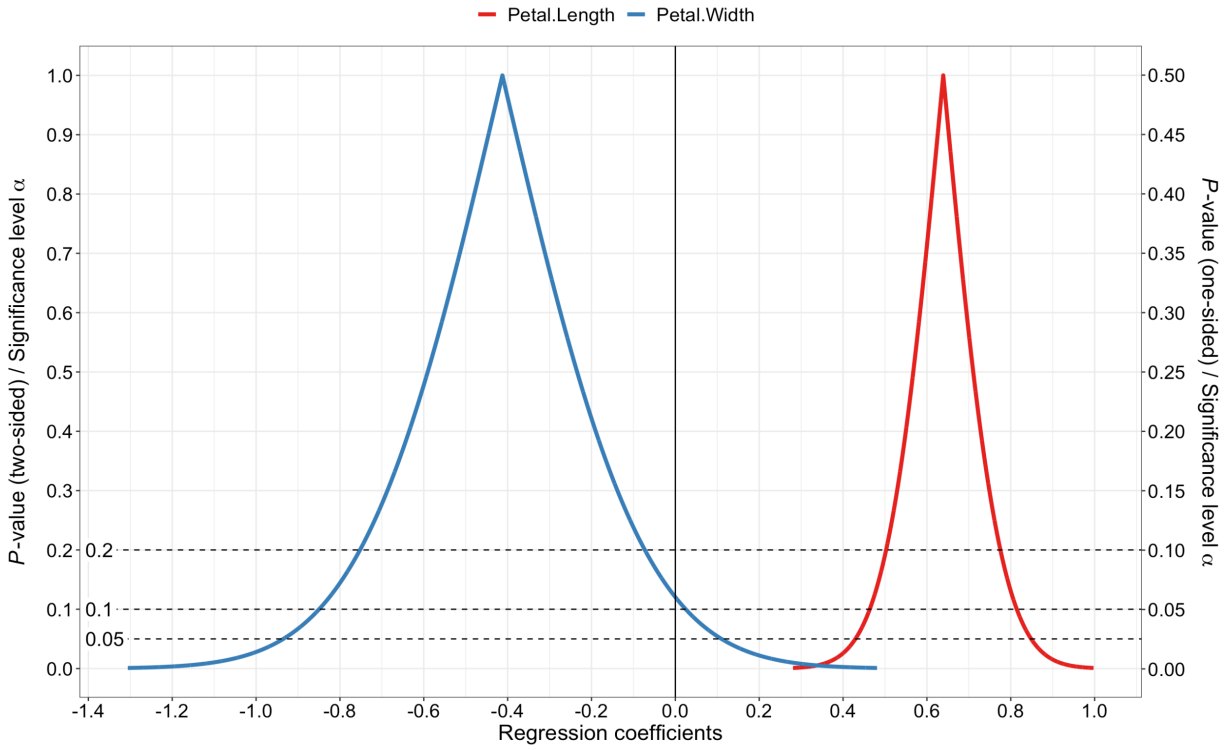
```r
#-------------------------------------------------------------------------------
# Le graphe (conf_dist) pvalue de Sepal.Width
#-------------------------------------------------------------------------------

res1 <- conf_dist(
  estimate = c(0.61196)
  , df = c(96 )
  , stderr = (0.31620)
  , type = "linreg"
  , plot_type = "p_val"
  , n_values = 1e4L
  , conf_level = c(0.95, 0.90, 0.80)
  , null_values = c(0)
  , trans = "identity"
  , alternative = "two_sided"
  , log_yaxis = TRUE
  , cut_logyaxis = 0.05
  , xlab = "Coefficient Petal.Width "
  , together = FALSE
  , plot_p_limit = 1 - 0.999
  , plot_counternull = FALSE
  , plot = TRUE)
```



Graphe PETAL.LENGTH et PETAL.WIDTH

Ici nous allons faire la même démarche mais cette fois-ci pour les coefficients associés aux deux variables Petal.Length et Petal.Width. D'après de le graphe ci-dessous, on constate que pour un seuil de significativité de 5% le coefficient associe Petal.Width est diffèrent de zero , par contre, le coefficient associé à la variable Petal.Length n'est pas diffèrent de zéro car sa p-value et supérieur de 5% de plus il a un intervalle de confiance qui contient le zéro.

```
res2 <- conf_dist(
  estimate = c(0.63897, -0.41249 )
  , df = c(96 , 96 )
  , stderr = c( 0.10560,  0.26325)
  , type = "linreg"
  , plot_type = "p_val"
  , n_values = 1e4L
  , est_names = c("Petal.Length", "Petal.Width")
  , conf_level = c(0.95, 0.90, 0.80)
  , null_values = c(0)
  , trans = "identity"
  , alternative = "two_sided"
  , log_yaxis = FALSE
  , cut_logyaxis = 0.05
  , xlab = "Regression coefficients"
  , together = TRUE
  , same_color = FALSE
  , plot_p_limit = 1 - 0.999
  , plot_counternull = FALSE
  , inverted = FALSE)
```



Coefficient de corrélation de PEARSON

Dans l'exemple ci-dessous nous allons tester L'existence d'une relation linaire entre Sepal.Width et Petal.Length à travers le coefficient de corrélation de PEARSON. Le coefficient de Pearson est un indice reflétant une relation linéaire entre deux variables continues. cet indice varie entre- 1et 1 : 0 indique une relation nulle entre les deux variables, une valeur négative (corrélation négative) signifiant que lorsque l'une des variables augmente, l'autre diminue; tandis que la valeur positive (corrélation positive) indique que les deux variables varient ensemble dans le même sens. les hypothèses du test de corrélation de de PEARSON sont les suivantes:

H0: Pas de corrélation entre les deux variables : P = 0 VS H1: Corrélation entre les deux variables : P !=0

```
cor.test(iris$Sepal.Width, iris$Petal.Length, alternative = "two.sided", method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  iris$Sepal.Width and iris$Petal.Length
## t = -7.4763, df = 98, p-value = 3.265e-11
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.7145027 -0.4607906
## sample estimates:
##        cor
## -0.6026631
```

Interprétation des Résultats du coefficient de corrélation de PEARSON

la corrélation observée dans cet échantillon entre le Petal.Length et le Sepal.Width est de -0.6026.

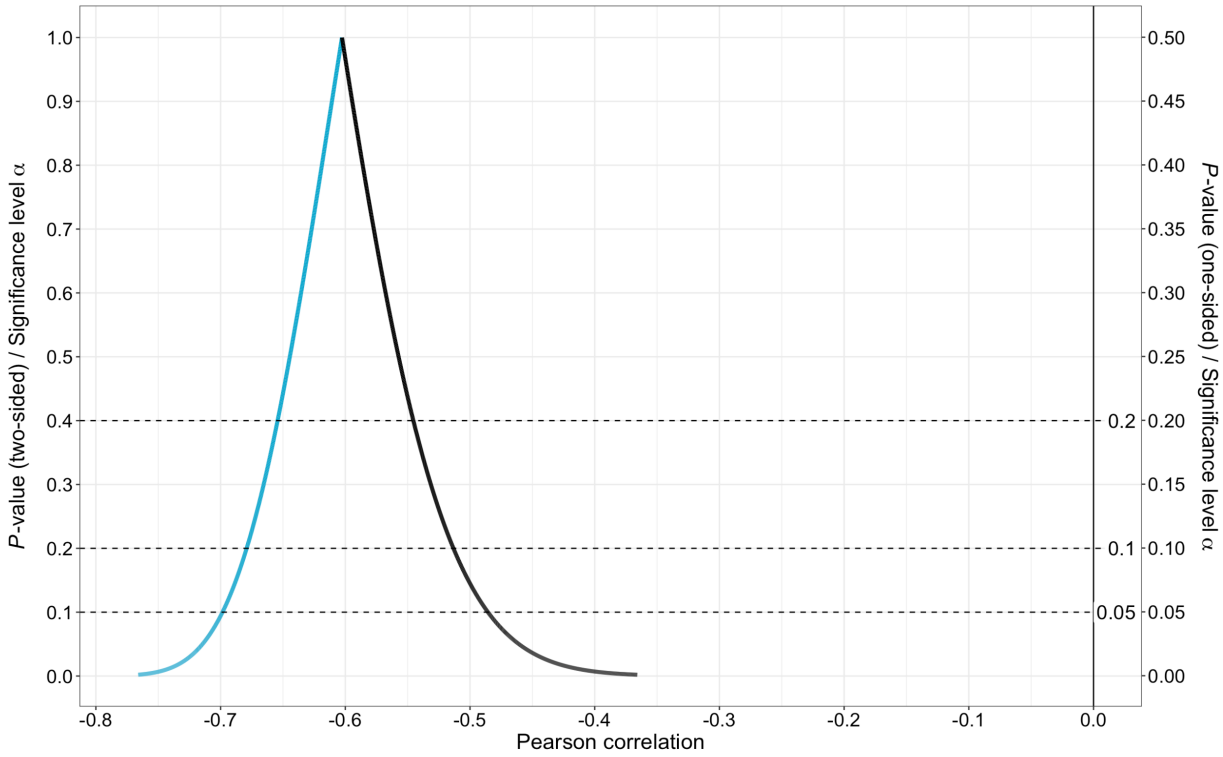
La probabilité d'avoir une corrélation aussi élevée dans cet échantillon est de 3.265e-11.

Etant donné que cette probabilité est faible (inférieure au seuil de significativité = 0.05), on peut rejeter H0 et conclure

que la corrélation entre le Petal.Length et le Sepal.Width est significativement négative.

Graphe

```
res3 <- conf_dist(
  estimate = c(-0.6026631)
  , n = 100
  , type = "pearson"
  , plot_type = "p_val"
  , n_values = 1e4L
  , conf_level = c(0.95, 0.90, 0.80)
  , null_values = c(0)
  , trans = "identity"
  , alternative = "one_sided"
  , log_yaxis = FALSE
  , cut_logyaxis = 0.05
  , xlab = "Pearson correlation"
  , together = TRUE
  , plot_p_limit = 1 - 0.999
  , plot_counternull = FALSE)
```



le graphique ci-dessus, nous montre que pour un seuil de significativité de 5% le coefficient de Pearson de la population possède un intervalle de confiance qui ne contient pas le zéro. D'où le rejet de l'hypothèses nulle.

Informations supplémentaires:

DATA IRIS

Le jeu de données Iris connu aussi sous le nom de Iris de Fisher est un jeu de données multivariées présenté en 1936 par Ronald Fisher dans son papier The use of multiple measurements in taxonomic problems comme un exemple d'application de l'analyse discriminante linéaire. Il est parfois aussi appelé Iris d'Anderson du nom d'Edgar Anderson qui a collecté ces données afin de quantifier les variations de morphologie des fleurs d'iris de trois espèces. Deux des trois espèces ont été collectées en Gaspésie. « Toutes sont du même champ, cueillies le même jour et mesurées le même jour par la même personne avec les mêmes outils de mesures. », Edgar Anderson (1935) Le jeu de données comprend 50 échantillons de chacune des trois espèces d'iris (Iris setosa, Iris virginica et Iris versicolor). Quatre caractéristiques ont été mesurées à partir de chaque échantillon : la longueur et la largeur des sépales et des pétales, en centimètres. Sur la base de la combinaison de ces quatre variables, Fisher a élaboré un modèle d'analyse discriminante linéaire permettant de distinguer les espèces les unes des autres.

Create And Plot P-Value Functions, S-Value Functions, Confidence Distributions And Confidence Densities Conf_dist

The function **conf_dist** generates confidence distributions (cdf), confidence densities(pdf), Shannon suprisal (s-value) functions and p-value functions for several commonly used estimates. In addition, counternulls (see Rosenthal et al. 1994), point estimates and the area under the confidence curve (AUCC) are calculated.

Details

P-value functions and confidence intervals are calculated based on the t-distribution for t-tests, linear regression coefficients, and gamma regression models (GLM). The normal distribution is used for logistic regression, poisson regression and cox regression models. For correlation coefficients, Fisher's transform is used using the corresponding variances (see Bonett et al. 2000). P-value functions and confidence intervals for variances are constructed using the Chi2 distribution. Finally, Wilson's score intervals are used for one proportion. For differences of proportions, the Wilson score interval with continuity correction is used (Newcombe 1998).

Value

conf_dist returns four data frames and if plot = TRUE was specified, a ggplot2-plot object: res_frame (contains parameter values (e.g. mean differences, odds ratios etc.), p-values (one- and two-sided), s-values, confidence distributions and densities, variable names and type of hypothesis), conf_frame (contains the specified confidence level(s) and the corresponding lower and upper limits as well as the corresponding variable name), counternull_frame (contains the counternull and the corresponding null values), point_est (contains the mean, median and mode point estimates) and if plot = TRUE was specified, aucc_frame contains the estimated AUCC (area under the confidence curves) calculated by trapezoidal integration on the untransformed scale, plot (a ggplot2 object).

Fonctions summaray(), lm()

summaray()

summary() permet de produire les sorties pour chaque regression présentées précédemment. Ainsi, cette fonction affiche : les coefficients estimés, leur écart-type, et la valeur de la statistique t de Student ainsi que la p-value (probabilité que le coefficient soit significativement différent de zéro) associées à chaque coefficient. Sont aussi présentés le R2 et R2 ajusté, ainsi que la statistique F de Fisher (testant la significativité globale des variables), son degré de liberté, et la p-value associée.

lm()

Jetons un coup d'œil la fonction lm() ("Linear Model") qui est plus moderne et plus générale. Pour cette fonction seul l'interface formule est possible pour spécifier des variables. On écrira alors par exemple:

- lm(mortinf ~urb, data=pays)
- lm(mortinf ~urb + tert, data=pays)
- lm(mortinf ~urb + tert+urb*tert, data=pays)

Les formules permettent d'utiliser une forme très naturelle d'équation pour spécifier le modèle que l'on veut faire estimer, y compris la spécification de termes d'interaction.

Le coefficient de corrélation

Définition:

Le test de corrélation permet d'étudier l'association (ou dépendance) entre deux ou plusieurs variables. Par exemple, lorsque l'on souhaite savoir s'il y a une association entre les poids des enfants et de leurs pères, le coefficient de corrélation peut être calculé pour répondre à cette question.
S'il n'y a aucun lien entre les deux variables (les poids des pères et des enfants), le poids moyen des enfants devrait être le même quelque soit le poids du père et vice versa.Il existe différentes méthodes pour l'analyse de corrélations : les tests de Pearson, de kendall et de Spearman.

Corrélation de Pearson:

La corrélation de Pearson mesure une dépendance linéaire entre deux variables (x et y). C'est une méthode dite paramétrique car elle dépend de la distribution des données. Cette méthode n'est conseillée que lorsque les variables suivent une loi normale. Dans le cas contraire, il faudrait utiliser les tests de corrélation non-paramétriques de type kendall et Spearman. Le graphique de y = f(x) est appelé droite de régression.

La formule de la corrélation de **Pearson** est : $r = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2(y_i-\overline{y})^2}}$

$\overline{x} et \overline{y}$ représentent les moyennes des variables x et y. La p-value (ou niveau de significativité) de la corrélation peut être déterminée:

- en utilisant la table des valeurs critiques de coefficient de corrélation pour un degré de liberté : dl=n2.
- ou en calculant la valeur t de Student $\frac{r}{\sqrt{1-r^2}}\sqrt{n-2}$

Dans ce cas la p-value correspondante est déterminée en utilisant la table de Student pour dl=n-2