Taller Estadística en R: Sesion 4

Abril 2022

Prueba T, ANOVAS, Correlaciones y Regresiones

Prueba T

La prueba T es una prueba que permite comparar las medias entre dos grupos. Es una prueba poderosa que funciona tanto con N grandes como pequeños.

La distribución t es aproximadamente normal, pero más ancha (las colas son más pesadas). El N mínimo para que t sea normal es 31.

Supuestos de la prueba t

- *Variables con distribución normal:* Este supuesto se puede violar si las distribuciones son parecidas
- *Homocedasticidad:* Si las desviaciones estándar no varían en más de 3X se puede violar, si varían en más de 3X se usa la prueba t aproximada de Welch

Hay 3 tipos esenciales de pruebas T: prueba t de una muestra, de dos muestras independientes o de muestras pareadas.

Prueba t de una muestra Se usa cuando solo hay una muestra y se está comparando contra una media establecida bajo la hipótesis nula. La fórmula es:

$$t = \frac{\bar{Y} - \mu}{SE_{\bar{Y}}}$$

$SE_{\bar{Y}}$ desviación estándar de la muestra y μ es la media esperada bajo la nula

$$SE_{\bar{Y}} = \frac{S}{\sqrt{n}}$$

Con t hallado, podemos rechazar o no la hipótesis nula al comparar con la tabla.Si el T hallado es mayor que el T crítica rechazamos la hipótesis nula. Los grados de libertad son N-1

En R se puede hacer con el comando **t.test()**

Por ejemplo digamos que queremos saber si la temperatura corporal de unos cangrejos es diferente de la temperatura ambiente.

H_0= La media de la temperatura corporal es igual a la temperatura ambiente H_A= La media de la temperatura corporal es diferente de la temperatura ambiente

```
t_ambiente<-24.3 #asumimos que la t ambiente es de 24.3
t_cangrejos<-c(rep(25.3,3),rep(23,5),rep(26,2),rep(24,10))
t.test(t_cangrejos,mu=t_ambiente,conf.level=0.95)
```

```
##
##  One Sample t-test
##
## data:  t_cangrejos
## t = -0.71921, df = 19, p-value = 0.4808
```

```
## alternative hypothesis: true mean is not equal to 24.3
## 95 percent confidence interval:
##  23.69393 24.59607
## sample estimates:
## mean of x
##    24.145
```

```
qt(0.19,df=19) #valor critico
```

```
## [1] -0.8988173
```

Como el valor T hallado dio menos que el valor critico (0.71<1.72), fallamos en rechazar la hipótesis nula y concluimos que no hay diferencia entre la temperatura de los cangrejos y la temperatura ambiente.

Prueba T pareada Esta se usa cuando las medidas entre grupos no son idependientes, es decir, cada muestra tiene los dos tratamientos.

La fórmula es parecida a la anterior, pero en vez de usar la media usamos la media de la diferencia

$$t = \frac{= \bar{d} - \mu_d}{SE_{\bar{d}}}$$

En este caso los grados de libertad son el número de pares-1

También se pueden hacer en R usando *t.test()*, pero hay que cambiar el argumento PAIRED

Por ejemplo queremos mirar si el número de especies cambia según si un río fue o no restaurado. Las muestras son pareadas, puesto que en cada parcela se tomó muestra para río restaurado y muestra para río sin restaurar.

```
resta<-c(12,23,45,10,27)
no_resta<-c(50,38,60,67,100)
t.test(resta,no_resta,paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  resta and no_resta
## t = -3.4525, df = 4, p-value = 0.026
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -71.445716  -7.754284
## sample estimates:
## mean of the differences
##                   -39.6
```

```
qt(0.95,df=4)
```

```
## [1] 2.131847
```

Como el valor T hallado dio más que el valor critico (2.13<3.45), rechazamos la hipótesis nula y concluimos que la riqueza de especies es diferentes entre ríos restaurados y ríos sin restaurar

Prueba T para muestras independientes Si las muestras son de dos grupos independientes hay que especificarlo. La hipótesis nula sigue siendo que la media entre grupos es igual.

La fórmula es:

$$t = \frac{\bar{Y_1} - \bar{Y_2}}{\frac{S}{\sqrt{n}}}$$

Y los grados de libertad son la suma de los grados de libertad de ambos grupos

En R se hace también como las anteriores

Digamos que queremos saber si el número de insectos en las lechugas difiere entre lechugas resistentes y lechugas no resistentes

```
resistentes<-c(8,11,10,10,10,10,10,11,11,10)
no_resistentes<-c(16,17,18,16,15,16,17,18,18,17)
var.test(resistentes,no_resistentes)
```

```
##
##  F test to compare two variances
##
## data:  resistentes and no_resistentes
## F = 0.71875, num df = 9, denom df = 9, p-value = 0.6307
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.1785273 2.8936833
## sample estimates:
## ratio of variances
##            0.71875
```

```
t.test(resistentes,no_resistentes,var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  resistentes and no_resistentes
## t = -15.648, df = 18, p-value = 6.34e-12
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -7.599561 -5.800439
## sample estimates:
## mean of x mean of y
##      10.1      16.8
```

```
#Como nuestras varianzas son iguales debemos añadir var.equal=TRUE

#Si no añadimos ese argumento, R hace una prueba t de Welch

qt(0.95,df=18) #Para mirar el valor T crítico
```

```
## [1] 1.734064
```

Como el t hallado es mayor que el tabulado, rechazamos la hipótesis nula y concluimos que el número de insectos difiere según el tipo de lechuga.

Análisis de varianza (ANOVA)

Un análisis de varianza permite comparar las medias de diferentes grupos o categorías.

La variable de entrada es una variable de tipo categórico y la de salida es de tipo numérico.

La hipótesis nula es que no hay diferencia entre las medias de las diferentes categorías y la hipótesis alterna es que *al menos* una de las medias es diferente

Supuestos del ANOVA

- Normalidad:
- Muestreo al azar
- Homocedasticidad
- Balanceo

ANOVA de 1 vía En este ANOVA hay una variable de entrada y una de salida.

Se puede hacer en R con el comando **aov()**

Digamos que queremos saber si hay diferencias entre el crecimiento de ciertos animales según el tipo de alimento

```
alimento<-c(rep("Alimento 1",5),rep("Alimento 2",5),
            rep("Alimento 3",4),rep("Alimento 4",5))
peso<-c(60.8,67,65,68.6,61.7,68.7,
        67.7,75,73.3,71.8,69.6,77.1,75.2,71.5,61.9,64.2,63.1,66.7,60.3)

tabla<-data.frame(alimento,peso)
anova<-aov(tabla$peso~tabla$alimento)
summary(anova)
```

```
##                Df Sum Sq Mean Sq F value   Pr(>F)
## tabla$alimento  3  338.9  112.98   12.04 0.000283 ***
## Residuals      15  140.8    9.38
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Como el valor F es mayor que el F de la tabla (12.04>3.28) rechazamos la hipótesis nula y concluimos que al menos una de las medias de los alimentos es diferente.

Prueba de Tukey-Kramer Si encontramos que hay diferencias entre medias, para saber cuál es la media diferente podemos hacer la prueba post-hoc de Tukey

Esto se puede hacer en R con el comando *TukeyHSD()*

Utilizando el ejemplo anterior vamos a hacer la prueba de Tukey

```
TukeyHSD(anova)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = tabla$peso ~ tabla$alimento)
##
## $`tabla$alimento`
##                          diff        lwr       upr     p adj
## Alimento 2-Alimento 1   6.68   1.096263 12.263737 0.0168421
## Alimento 3-Alimento 1   8.73   2.807553 14.652447 0.0034914
```

```
## Alimento 4-Alimento 1  -1.38  -6.963737  4.203737 0.8906642
## Alimento 3-Alimento 2   2.05  -3.872447  7.972447 0.7530266
## Alimento 4-Alimento 2  -8.06 -13.643737 -2.476263 0.0041505
## Alimento 4-Alimento 3 -10.11 -16.032447 -4.187553 0.0009497
```

Correlaciones

Las correlaciones sirven para medir la fuerza de la relación entre dos variables numéricas. Son un análisis parecido al de las regresiones, con la diferencia de que en las correlaciones no hay variable dependiente y variable independiente, sino que ambas variables se consideran independientes por lo que no se puede inferir causalidad a partir de la correlación.

Hay diferentes tipos de correlaciones que se usan en diferentes escenarios:

Correlación de Pearson (r^2)

La correlación de Pearson ayuda a medir la fuerza de la asociación entre dos variables. La hipótesis nula más comúnmente probada con el coeficiente de correlación de Pearson es que es igual a cero, es decir, el coeficiente de correlación de la población es igual a cero y no existe una relación lineal entre las dos variables de la población.

Supuestos

- Muestreo aleatorio
- Independencia de las observaciones
- La distribución de probabilidad conjunta de Y1 e Y2 es bivariada normal. Si una o ambas variables tienen distribuciones no normales, entonces su distribución conjunta no puede ser bivariante normal y cualquier relación entre las dos variables podría no ser lineal, por lo que es importante verificar una relación no lineal con un diagrama de dispersión simple y distribuciones asimétricas de las variables con diagramas de caja.

Correlación de Spearman (r_s)

Utilice la correlación de rango de Spearman para probar la asociación entre dos variables clasificadas, o una variable clasificada y una variable de medición. También puede usar la correlación de rango de Spearman en lugar de la regresión/correlación lineal para dos variables de medición si le preocupa la falta de normalidad.

Correlación de Kendall (τ)

La correlación de Kendall prueba la relación entre dos variables en ausencia de normalidad. Tanto la correlación de Spearmann como la de Kendall son medidas más conservadoras que la correlación de Pearson cuando se cumplen los supuestos de distribución. Tenga en cuenta que estos análisis de correlación no paramétricos no detectan todas las asociaciones no lineales entre variables, solo relaciones monótonas.

Los tres tipos de correlaciones se pueden hacer en R usando el comando **cor()** o **cor.test()**, lo que cambia es el argumento *method*, que determina el tipo de correlación

Digamos que queremos analizar si el peso de unos murciélagos está correlacionado con su largo total

```
peso <- c(51, 59, 49, 54, 50, 55, 48, 53, 52, 57)
largo<-c(3.35, 3.8, 3.2, 3.75, 3.15, 3.3, 3.1, 3.65, 3.4, 3.5)
shapiro.test(peso)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  peso
## W = 0.97343, p-value = 0.9207
```

```
shapiro.test(largo)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  largo
## W = 0.93757, p-value = 0.5263
```

```
#como ambas son normales podemos hacer la prueba de Pearson
cor(largo,peso,method="pearson") #esta prueba solo les da el R2
```

```
## [1] 0.7794691
```

```
cor.test(largo,peso,method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  largo and peso
## t = 3.5194, df = 8, p-value = 0.007853
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2942560 0.9452104
## sample estimates:
##       cor
## 0.7794691
```

Si no tuvieramos datos normales deberíamos hacer la de Spearman y si tenemos rangos la de Kendall.

Regresión lineal simple

Utilice la regresión lineal cuando desee saber si una variable de medición está asociada con otra variable de medición; desea medir la fuerza de la asociación (r^2); o desea una ecuación que describa la relación y pueda usarse para predecir valores desconocidos.

En este caso, a diferencia de la correlación, se puede establecer una relación de causalidad y hay una variable independiente y una variable dependiente.

La hipótesis nula de esta prueba es que la pendiente de la línea de mejor ajuste es igual a cero; en otras palabras, a medida que la variable X aumenta, la variable Y asociada no aumenta ni disminuye.

Supuestos de la regresión lineal simple

- *Normalidad y homocedasticidad:* Dos suposiciones, similares a las de ANOVA, son que para cualquier valor de X, los valores de Y se distribuirán normalmente y serán homocedásticos. Aunque rara vez tendrá suficientes datos para probar estas suposiciones, a menudo se violan. Afortunadamente, numerosos estudios de simulación han demostrado que la regresión y la correlación son bastante resistentes a las desviaciones de la normalidad; esto significa que incluso si una o ambas variables no son normales, el valor de P será inferior a 0,05 aproximadamente el 5% de las veces si la hipótesis nula es verdadera (Edgell y Noon 1984, y referencias allí). Entonces, en general, puede usar la regresión/correlación lineal sin preocuparse por la no normalidad.

-*Linearidad:* La regresión lineal supone que los datos se ajustan a una línea recta. Si observa los datos y la relación parece curva, puede probar diferentes transformaciones de datos de la X, la Y o ambas, y ver cuál hace que la relación sea recta.

-Independencia

En R, se pueden hacer regresiones lineales usando el comando *lm()* en donde los argumentos son el modelo (variable dependiente~variable independiente) y los datos

Digamos que tenemos un set de datos con registros para múltiples sitios de las variables ambientales y de la riqueza de especies y queremos saber si esa riqueza de especie es causada por las variables ambientales

```
datos_sitios<-read.table("datossitios.txt",header = TRUE,sep = '\t')
summary(lm(riqueza~altitud,data=datos_sitios))
```

```
##
## Call:
## lm(formula = riqueza ~ altitud, data = datos_sitios)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.808 -4.427 -1.383  3.156 27.154
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.3056     0.2734   37.70   <2e-16 ***
## altitud       0.6103     0.2737    2.23   0.0263 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.682 on 430 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.01143,    Adjusted R-squared:  0.00913
## F-statistic: 4.971 on 1 and 430 DF,  p-value: 0.02628
```

El valor P nos muestra que sí hay una relación entre la riqueza y la altitud del sitio.

Regresión múltiple

Una extensión común de la regresión lineal simple es el caso en el que hemos registrado más de una variable predictora. Cuando todas las variables predictoras son continuas, los modelos se denominan modelos de regresión múltiple.

La hipótesis nula básica que podemos probar cuando ajustamos un modelo de regresión lineal múltiple es que todas las pendientes de la regresión parcial son iguales a cero.

Supuestos de la prueba:

- Normalidad
- Homocedasticidad
- Linealidad
- **No colinealidad:** En los modelos lineales múltiples los predictores deben ser independientes, no debe de haber colinialidad entre ellos. La colinialidad ocurre cuando un predictor está linealmente relacionado con uno o varios de los otros predictores del modelo o cuando es la combinación lineal de otros predictores. Como consecuencia de la colinialidad no se puede identificar de forma precisa el efecto individual que tiene cada una de las variables colineales sobre la variable respuesta, lo que se traduce en un incremento de la varianza de los coeficientes de regresión estimados hasta el punto que resulta prácticamente imposible establecer su significancia estadística. Además, pequeños cambios en los datos provocan grandes cambios en las estimaciones de los coeficientes.

No existe un método estadístico concreto para determinar la existencia de colinialidad o multicolinialidad entre los predictores de un modelo de regresión, sin embargo, se han desarrollado numerosas reglas prácticas que tratan de determinar en qué medida afecta a la estimación y contraste de un modelo. Los pasos recomendados a seguir son:

Si el coeficiente de determinación R2 es alto pero ninguno de los predictores resulta significativo, hay indicios de colinialidad.

Generar un modelo de regresión lineal simple entre cada uno de los predictores frente al resto. Si en alguno de los modelos el coeficiente de determinación R2 es alto, estaría señalando a una posible colinialidad.

Cuando se intenta establecer relaciones causa-efecto, la colinialidad puede llevar a conclusiones muy erróneas, haciendo creer que una variable es la causa cuando en realidad es otra la que está influenciando sobre ese predictor.

En este caso, podemos hacer una regresión múltiple usando los datos que tenemos.

```
datos_sitios<-na.omit(datos_sitios[,(-1)])
round(cor(datos_sitios),2)
```

```
##                   altitud precipitacion temperatura est.temperatura
## altitud              1.00         -0.28       -0.99           -0.47
## precipitacion       -0.28          1.00        0.28           -0.23
## temperatura         -0.99          0.28        1.00            0.40
## est.temperatura     -0.47         -0.23        0.40            1.00
## est.precipitacion   -0.40         -0.46        0.36            0.58
## riqueza              0.11          0.32       -0.11           -0.33
##                   est.precipitacion riqueza
## altitud                       -0.40    0.11
## precipitacion                 -0.46    0.32
## temperatura                    0.36   -0.11
## est.temperatura                0.58   -0.33
## est.precipitacion              1.00   -0.34
## riqueza                       -0.34    1.00
```

```
#como hay poca correlación entre variables, podemos usarlas

modelo1<-lm(riqueza~altitud+precipitacion+temperatura+
              est.temperatura+est.precipitacion,data = datos_sitios)
summary(modelo1)
```

```
##
## Call:
## lm(formula = riqueza ~ altitud + precipitacion + temperatura +
##     est.temperatura + est.precipitacion, data = datos_sitios)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.4243  -3.3791  -0.9367   2.3415  22.7924
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        10.3056     0.2464  41.822  < 2e-16 ***
## altitud            -6.7018     1.8668  -3.590 0.000369 ***
## precipitacion       0.9943     0.3543   2.806 0.005247 **
## temperatura        -6.5617     1.7352  -3.782 0.000178 ***
## est.temperatura    -1.5904     0.3508  -4.533 7.55e-06 ***
## est.precipitacion  -0.9235     0.3808  -2.425 0.015703 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.122 on 426 degrees of freedom
```

```
## Multiple R-squared:  0.2043, Adjusted R-squared:  0.195
## F-statistic: 21.88 on 5 and 426 DF,  p-value: < 2.2e-16
```

```
#Podemos hacer varios modelos incluyendo diferentes variables

modelo2<-lm(riqueza~altitud+precipitacion+temperatura,data = datos_sitios)

summary(modelo2)
```

```
##
## Call:
## lm(formula = riqueza ~ altitud + precipitacion + temperatura,
##     data = datos_sitios)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.351  -3.709  -1.273   2.837  23.471
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    10.3056     0.2549  40.432  < 2e-16 ***
## altitud        -0.6070     1.5722  -0.386    0.700
## precipitacion   2.1611     0.2661   8.122 4.95e-15 ***
## temperatura    -1.8484     1.5735  -1.175    0.241
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.298 on 428 degrees of freedom
## Multiple R-squared:  0.1447, Adjusted R-squared:  0.1387
## F-statistic: 24.13 on 3 and 428 DF,  p-value: 1.908e-14
```

```
modelo3<-lm(riqueza~precipitacion+temperatura,data = datos_sitios)
summary(modelo3)
```

```
##
## Call:
## lm(formula = riqueza ~ precipitacion + temperatura, data = datos_sitios)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.242  -3.691  -1.274   2.859  23.563
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    10.3056     0.2546  40.473  < 2e-16 ***
## precipitacion   2.1619     0.2658   8.133 4.56e-15 ***
## temperatura    -1.2496     0.2658  -4.701 3.49e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.292 on 429 degrees of freedom
## Multiple R-squared:  0.1444, Adjusted R-squared:  0.1404
## F-statistic: 36.19 on 2 and 429 DF,  p-value: 2.983e-15
```

```
modelo4<-lm(riqueza~altitud+precipitacion+temperatura+est.precipitacion,data = datos_sitios)
summary(modelo4)
```

```
##
## Call:
## lm(formula = riqueza ~ altitud + precipitacion + temperatura +
##     est.precipitacion, data = datos_sitios)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.311  -3.665  -1.060   2.618  22.884
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        10.3056     0.2520  40.896  < 2e-16 ***
## altitud            -3.0628     1.7235  -1.777  0.07628 .
## precipitacion       1.3953     0.3509   3.976 8.21e-05 ***
## temperatura        -3.6072     1.6445  -2.193  0.02882 *
## est.precipitacion  -1.2596     0.3819  -3.298  0.00106 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.238 on 427 degrees of freedom
## Multiple R-squared:  0.1659, Adjusted R-squared:  0.1581
## F-statistic: 21.24 on 4 and 427 DF,  p-value: 5.485e-16
```

```
#Podemos comparar qué modelo es mejor con el AIC

AIC(modelo1)
```

```
## [1] 2645.24
```

```
AIC(modelo2)
```

```
## [1] 2672.462
```

```
AIC(modelo3)
```

```
## [1] 2670.612
```

```
AIC(modelo4)
```

```
## [1] 2663.595
```

De estos resultados podemos decir que el mejor modelo es aquel que utiliza todas las variables porque tiene menor AIC y este modelo nos dice que si bien todas las variables son importantes, las más significativas son temperatura y altitud, por lo cual podemos concluir que la riqueza de especie se asocia mucho a la temperatura y la altitud.