

Estadística Aplicada

Condiciones para pruebas de hipótesis

Clase 10

Ingeniería en ciberseguridad

La excelencia no se improvisa



1. INTRODUCCIÓN DE LA CLASE

En la clase anterior, habíamos visto los distintos usos de la prueba t.test. Habíamos mencionado, entonces, que esta prueba está relacionada con la distribución t y que existían algunos requisitos para que funcione adecuadamente. Específicamente, el primer requisito es que la muestra sobre la que estamos trabajando tenga una distribución normal o cuasi-normal. Por otro lado, indicamos que cuando t.test se usa para comparar si dos muestras, denotadas como X/Y, provienen de la misma población, los datasets deben tener varianzas iguales o al menos similares. En la presente clase, presentaremos algunas funciones que nos sirven para validar estos supuestos.

Clase 10:

Analizar información contextual sobre hábitat, infraestructura y movilidad, mediante técnicas estadísticas descriptivas e inferenciales, para la adecuada toma de decisiones

Reto # 4

10.) Condiciones para pruebas de hipótesis

Para que t.test funcione correctamente, es necesario que se cumplan dos condiciones: la distribución debe ser normal y la varianza, en el caso de comparación de dos datasets, debe ser similar.

10.1) Prueba de Shapiro-Wilk

La prueba de normalidad de Shapiro-Wilk es una de las tres pruebas generales de normalidad diseñadas para detectar las desviaciones de la normalidad. El test de Shapiro-Wilk plantea la hipótesis nula de que la muestra proviene de una distribución normal. Para la prueba, elegimos un nivel de significancia; la prueba permite escoger el nivel de significancia deseado, siendo los más comunes 0.05 o 0.01, y tenemos una hipótesis alternativa que sostiene que la distribución no es normal.

H_0 = La distribución es normal

H_1 = La distribución no es normal,

o más formalmente:

$H_0: X \sim N(\mu, \sigma^2)$

$H_1: X \not\sim N(\mu, \sigma^2)$

Esta prueba retorna un p-value. El valor p es el nivel más bajo de significancia al cual se puede rechazar la hipótesis nula; por lo tanto, se garantiza que la distribución es normal si el p-value es mucho mayor que 0.05.

Esta prueba también realiza una prueba no paramétrica medida por el estadístico W. Este debe tener un valor aproximadamente igual a 1 para muestras de distribuciones normales. Los valores grandes indican no normalidad; más específicamente, el valor W debe compararse con un W-crítico, mismo que varía de acuerdo con el tamaño de la muestra y el nivel de confianza.

Valores referenciales figuran en la siguiente tabla.

Tabla 1

Valores del W-crítico en función de la cantidad de observaciones

Number or Samples	W(critical)
3	0.767
4	0.748
5	0.762
6	0.788
7	0.803
8	0.818
9	0.829
10	0.842
11	0.850
12	0.859
13	0.866
14	0.874

Para el cálculo del $W_{critical}$, no podemos utilizar las funciones $qnorm$ o qt , dado que no podemos asegurar su distribución. En su lugar, usaremos una distribución gamma. Esta distribución de la familia de probabilidad continua contiene dos parámetros: parámetro de forma K y de escala θ . El siguiente código ayuda a encontrar el $W_{critical}$ para cualquier combinación de número de observaciones y nivel de confianza.

```
#Cálculo de Wcritico

tmpf <- function(gshape=20,n=50) {
  shapiro.test(qgamma((1:n)/(n+1),scale=1,shape=gshape))}

# Encontrar el parámetro de forma que corresponde a un valor p particular

find.shape <- function(n,alpha) {
  uniroot(function(x) tmpf(x,n)$p.value-alpha,
    interval=c(0.01,100))$root}

find.W <- function(n,alpha) {
  s <- find.shape(n,alpha)
  tmpf(s,n=n)$statistic
}

#Wcritico para 14 observaciones y alfa 0.05

find.W(14,0.05)

0.8753995

#Wcritico para 14 observaciones y alfa 0.0

find.W(14,0.01)

0.8238614
```

Distribución gamma

Describe el uso de las funciones gamma

[Enlace](#)

Veamos un ejemplo teórico

```
#En primer lugar creamos una distribución normal
vnormal <- rnorm(100, mean = 5, sd = 3)
shapiro.test(vnormal )

Shapiro-Wilk normality test

data:  rnorm(100, mean = 5, sd = 3)
W = 1, p-value = 0.2

#Dado que p-value es > 0.05 , la distribución del vector es normal

#El siguiente vector tiene una distribución uniforme
runif <- runif(100, min = 2, max = 4)
shapiro.test(vrunif)

shapiro.test(runif(100, min = 2, max = 4))

Shapiro-Wilk normality test

data:  runif(100, min = 2, max = 4)
W = 1, p-value = 0.0009

#Dado que runif nos provee un vector con distribución uniforme, el resultado
#es un pvalue mucho menor que 0.05
```

Veamos un ejemplo con el dataset “housing”, queremos entender si la distribución de los pies cuadrados de las construcciones sigue una distribución normal

```
read.csv("http://www.jaredlander.com/data/housing.csv", header=TRUE, sep=",")
housing<- na.omit(housing)
#Entendamos el dataset
names(housing)
#Nombre largos o con puntos es mejor cambiarlos
nombres <- names(housing)
names(housing) <- c("sector", "class", "TUnits", "Year", "GSqFt",
                    "Est_GIncome", "GIncomexSqFt", "Est_Expense",
                    "ExpSqFt", "NetIncome",
                    "FullValue", "valxSqFt", "Boro")
#Nos interesa saber si de todo el dataset el tamaño de los áreas siguen
#una distribución normal
shapiro.test(housing$GSqFt)
Shapiro-Wilk normality test
data: housing$GSqFt
W = 0.4567, p-value < 2.2e-16
#Con un pvalue menor que 0.05 y W menor que 1 concluimos que no sigue
Distribución normal
```

Si un conjunto de datos no se distribuye normalmente, podemos realizar una de las siguientes transformaciones para normalizarlo:

1. Transformación logarítmica: transforma el vector a $\log(\text{vector})$.
2. Transformación de raíz cuadrada: transforma el vector y a $\sqrt{\text{vector}}$

Al realizar estas transformaciones, la variable de respuesta suele acercarse a una distribución normal.

10.2) Prueba de Ansari-Bradley

Esta prueba se usa para probar la hipótesis nula (H_0) de que las varianzas de dos datasets, correspondientes a dos muestras, son idénticas, mientras que la hipótesis alternativa (H_1) indica que difieren por dispersión (algunos autores lo llaman escala). La prueba de Ansari también retorna un p-value que debe ser mucho mayor que 0.05 para asegurar que las varianzas son similares.

Figura 1

Función programática Ansari

```
ansari.test(x, ...)  
  
# S3 method for default  
ansari.test(x, y,  
            alternative = c("two.sided", "less", "greater"),  
            exact = NULL, conf.int = FALSE, conf.level = 0.95,  
            ...)
```

Veamos un ejemplo teórico

```
#Creamos un vector normal  
V1 <- rnorm(100)  
  
#Creamos un segundo vector normal pero con desviación estándar de 2  
V2 <- rnorm(100, 0,2)  
  
ansari.test(V1, V2)  
  
Ansari-Bradley test  
data: rnorm(100) and rnorm(100, 0, 2)  
AB = 5872, p-value = 5.892e-05  
alternative hypothesis: true ratio of scales is not equal to 1  
95 percent confidence interval:  
0.4035154 0.7085887  
sample estimates:  
ratio of scales  
0.5356437  
  
#Pvalue obtenido es menor que 0.05 indicando que la función detecta que las varianzas  
son diferentes.
```

De la figura 1 se puede ver claramente que la función solo permite comparar dos datasets; sin embargo, comúnmente nos topamos con situaciones en las cuales tenemos una variable de tipo factor con múltiples niveles. En estos casos, necesitamos garantizar que los valores de cada uno de estos niveles tengan una varianza similar. La pregunta que nos hacemos es: ¿Cómo puedo comparar múltiples datasets? La respuesta es ANOVA.

10.3) Pruebas de anova de 1 vía

Estadístico Fratio

Para entender cómo funciona anova, necesitamos conocer previamente sobre la distribución F. De acuerdo con Berman H,() “la distribución F es la distribución de probabilidad asociada con el estadístico F”. De la misma manera que la distribución Student T está relacionada con el estadístico t, la distribución F está asociada con el estadístico F (o Fratio).

El estadístico Fratio sirve para comparar varianzas de dos poblaciones; la comparación se hace en base a la razón (división) de las mismas. Sin embargo, debemos entender que los estadísticos de las poblaciones generalmente no son conocidos. Cuando se comparan las varianzas de dos poblaciones, se toma una muestra de cada población. Las varianzas de la muestra sirven como estimaciones de sus varianzas poblacionales respectivas. Una distribución F se forma por la razón de estas dos varianzas muestrales.

Debemos conocer también que existen 2 tipos de pruebas en las que usamos este estadístico. La prueba conocida como anova de 1 vía (aov), tiene por interés contestar la siguiente pregunta: “¿La varianza entre las medias de dos poblaciones es significativamente diferente?”

Análisis de variación usando anova

Describe los casos de usos y resultados

[Enlace](#)

La distribución F tiene la particularidad de que la relación de las varianzas se espera que sea mayor que 1, por lo tanto, al hacer la relación, colocaremos la varianza mayor en el numerador y la varianza menor en el denominador.

Para calcular el estadístico F, seleccionamos una muestra aleatoria de tamaño n_1 de la población normal 1 bajo análisis y que tenga una desviación estándar igual a S_1 .

Luego seleccionamos otra muestra aleatoria independiente de tamaño n_2 de otra población normal, que tenga una desviación estándar igual a S_2 .

Quedando su fórmula como se menciona en la siguiente figura

Figura 2

Razón F para comparar varianzas

$$F = \frac{S_1^2}{S_2^2}$$

Distribución F

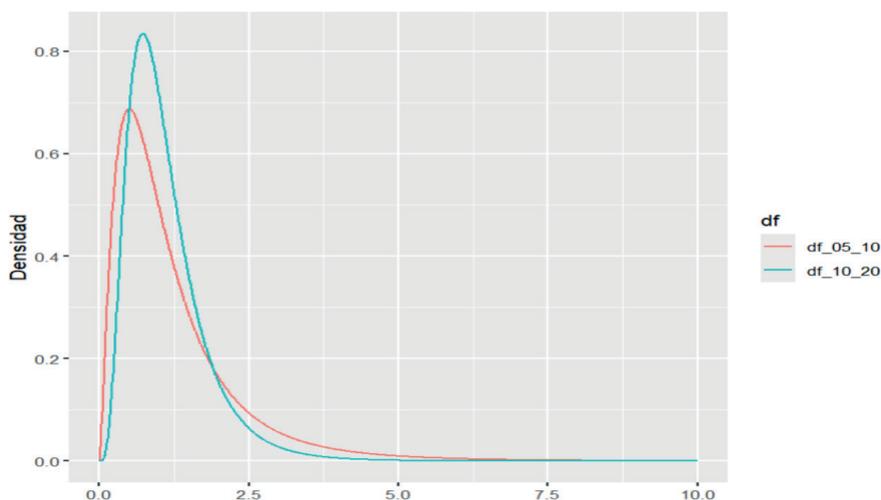
La distribución de todos los valores posibles del estadístico f se denomina distribución F , siendo $v_1 = n_1 - 1$ y $v_2 = n_2 - 1 = a$ grados de libertad de los dos sets bajo comparación.

La curva de la distribución F depende de los grados de libertad, v_1 y v_2 . Al describir una distribución F , el número de grados de libertad asociados con la desviación estándar en el numerador del estadístico f siempre se indica primero. Por lo tanto, $f(5, 10)$ se referiría a una distribución F con $v_1 = 5$ y $v_2 = 10$ grados de libertad; mientras que $f(10, 20)$ se referiría a una distribución F con $v_1 = 9$ y $v_2 = 5$ grados de libertad. Nótese que la curva representada por $f(5, 10)$ sería diferente de la curva representada por $f(10, 20)$.

Figura 3

Dos ejemplos de distribución F el primero con 10 y 20 grados de libertad y el segundo con 5 y 10 grados de libertad

Creación de autor: Alfonso Prado



Al igual que cuando hacemos una prueba de hipótesis tenemos una área de rechazo de H_0 y un área de aceptación de H_0 , en la distribución F vamos a tener una área de rechazo y una de no rechazo. Sin embargo, debido a que la razón F está restringida para exceder siempre de 1, resulta que solo tienen una zona de rechazo en la cola derecha. Por lo tanto, únicamente la mitad del área bajo la distribución es accesible como zona de rechazo y es necesario dividir entre 2 el valor de α seleccionado e identificar una zona de rechazo única en la cola derecha.

La delimitación de estas áreas estará dada por el intervalo de confianza con el que se desea trabajar, así por los grados de libertad del numerador y denominador de los sets bajo comparación. Para encontrar este valor, se utiliza la función $qf()$ cuya sintaxis se muestra en la figura xx

En los siguientes ejemplos, puede observarse cómo el valor correspondiente a alfa 0.05 y 0.01 varía con respecto a los grados de libertad del numerador y denominador.

```
qf(.95, 5, 10)
```

```
[1] 3.33
```

```
> qf(.95,10, 5)
```

```
[1] 4.74
```

```
> qf(.99, 5,10)
```

```
[1] 5.64
```

```
qf(.99,10,5)
```

```
[1] 10.1
```

Figura 4

Sintaxis de la función qf

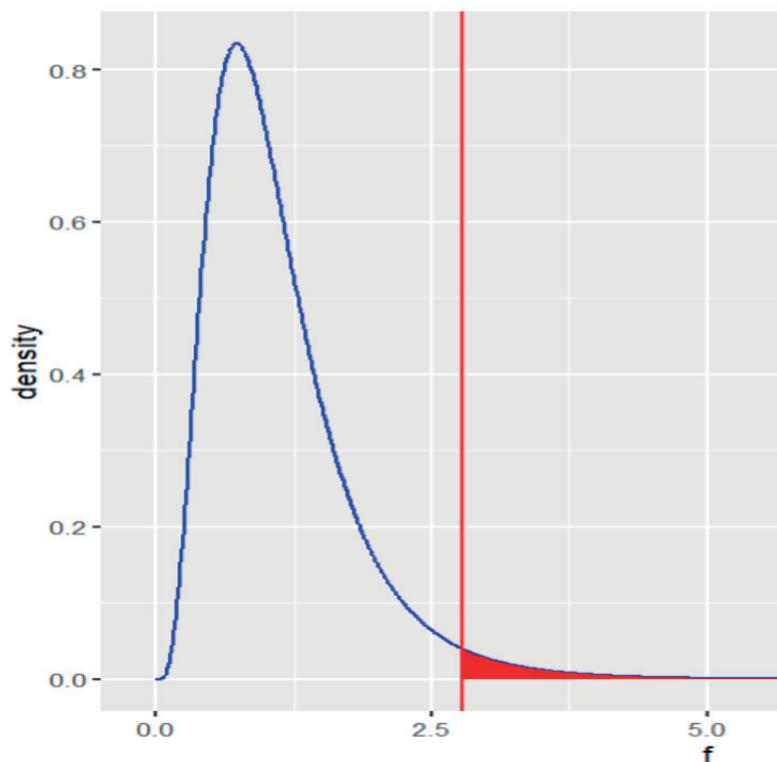
```
qf(p, df1, df2, ncp, lower.tail = TRUE, log.p = FALSE)
```

El valor retornado por la función qf() limitaría las zonas de rechazo como se muestra a continuación.

Figura 5

Zona de rechazo de una distribución F

Creación de autor Alfonso Prado



Funcionamiento de aov

La técnica de análisis de varianza (aov), también conocida como análisis factorial, constituye la herramienta básica para el estudio del efecto de uno o más factores (cada uno con dos o más niveles) sobre la media de una variable continua. Es, por lo tanto, la prueba estadística adecuada cuando se desea comparar las medias de dos o más grupos. Esta técnica también puede generalizarse para estudiar los posibles efectos de los factores sobre la varianza de una variable.

aov tiene las siguientes hipótesis y supuestos:

- ✓ H0: La media de todos los grupos considerados son iguales.
- ✓ H1: La media de todos los grupos no son todas iguales.
- ✓ En realidad, no indica cuál(es) son diferentes o cuál(es) son iguales, lo único que indica es que no son todos iguales.
- ✓ Pruebas adicionales son necesarias para establecer cuáles son iguales o no.
- ✓ Asume que los dataset son distribuidos normalmente y que sus desviaciones estándar son la similares.
- ✓ Los datos atípicamente extremos pueden invalidar por completo las conclusiones de un ANOVA. Si se observan residuos extremos hay que estudiar con detalle a que observaciones pertenecen, siendo aconsejable recalculer el ANOVA sin ellas y comparar los resultados obtenidos.

Función Programática

Como indicamos anteriormente, existen 2 funciones ANOVA que sirven para propósitos diferentes:

- ✓ aov es una variante donde los grupos se hallan en un mismo dataset . A veces llamado “1-way-ANOVA” los grupos se forman por los niveles de una variable tipo factor
- ✓ anova es similar, pero está diseñada para permitir comparar la varianza de dos modelos predictivos (Veremos esto en la próxima clase)

Figura 6
Sintaxis de la función aov

```
aov(formula, data = NULL, projections = FALSE, qr = TRUE,  
     contrasts = NULL, ...)
```

De acuerdo con Rodrigo (2017) “El estadístico estudiado en el ANOVA, conocido como *Fratio*, es la ratio entre la varianza de las medias de los grupos y el promedio de la varianza dentro de los grupos”. Sin embargo, como en todas las pruebas, los estadísticos per-se no indica mayor cosa, los estadísticos deben ser comparados con un valor crítico que marca la frontera entre zonas de rechazo y zona de no rechazo de H0.

En el caso de aov() el estadístico F debe ser comparado con un F-critico (obtenido de una distribución F) que depende de algunos parámetros (Intervalo de confianza, grados de libertad del numerador , grados de libertad del denominador).

Ejemplo con el dataset de construcciones

```
#El dataset contiene una variable categórica llamada class que define el tipo de construcción. Queremos determinar si la varianza en precios es igual para cada tipo de construcción
```

```
levels(housing$class)
```

```
[1] "R2-CONDOMINIUM" "R4-CONDOMINIUM" "R9-CONDOMINIUM"  
"RR-CONDOMINIUM"
```

```
aov_class <- aov(valxSqFt~class, data= housing)
```

```
summary(aov_class)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
class	3	1243078	414359	97.3	<2e-16 ***
Residuals	2526	10756882	4258		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

```
□
```

```
#Obtenemos el fcritico para los grados de libertad mencionados
```

```
qf(.95,3,2526)
```

```
[1] 2.608426
```

```
#El estadístico F reportado por anova es muy superior al fcritico ,por lo tanto indica que hay diferencias en las varianzas de los grupos de tipos de construcción
```

```
#Una segunda prueba con respecto a los vecindarios
```

```
levels(housing$Boro)
```

Referencias citadas en la Clase 10.

Berman H.G., “*F Distribution*”, descargado de <https://stattrek.com/probability-distributions/f-distribution>

Rodrigo J. (2017), ANOVA análisis de varianza para comparar múltiples medias, descargado de https://cienciadedatos.net/documentos/19_anova

Definición de los términos citados en la Clase 10.

Distribución gamma	La distribución de gamma es una distribución de probabilidad continua que se utiliza para modelar el tiempo de espera hasta que ocurran un número específico de eventos. Es una distribución de dos parámetros, lo que significa que se requiere dos parámetros para definir completamente la distribución. Los dos parámetros son el parámetro de forma, denotado por alfa (α) y el parámetro de escala, denotado por beta (β). El parámetro de forma controla la forma de la distribución, mientras que el parámetro de escala controla la propagación de la distribución.
Factores y niveles	<p>Un factor es una variable categórica con un número finito de valores o niveles. En R, los factores se utilizan habitualmente para realizar clasificaciones de los datos, estableciendo su pertenencia a los grupos o categorías determinados por los niveles del factor.</p> <p>Los niveles de un factor pueden estar codificados como valores numéricos o como caracteres. Independientemente del tipo, internamente R almacena los factores como números enteros, lo cual permite optimizar el uso de memoria.</p>

Profundización Clase 10.

Ansari

Describe el uso de la función con ejemplos

[Enlace](#)

Anova

Describe el uso de la función con ejemplos

[Enlace](#)



La excelencia no se improvisa

síguenos

