Análisis de la potencia de la prueba ante la violación del supuesto de normalidad en un modelo lineal

Juan José Jaikel, Manrique Chacón, Paula Rodríguez
Escuela de Estadística, Universidad de Costa Rica
juan.jaikel@ucr.ac.cr, man.c.r@hotmail.com, prm1004@gmail.com

RESUMEN

En la teoría se utilizan los modelos lineales asumiendo que los errores son normales. Este no es siempre el caso en la vida real por lo que se quiere observar qué sucede con cuando se inclumple este supuesto, es decir, cuando no se obtienen observaciones de la variable respuesta con distribución normal, sino que para términos de la simulación sean de distribución uniforme, log-normal y exponencial. Por tal motivo se desea simular cuatro modelos con tres tratamientos cada uno, con el propósito de extraer y analizar la potencia de la prueba para la diferencia de medias. Para las pruebas se establece un nivel de significancia de 5%.

PALABRAS CLAVE: Potencia, diferencia de medias, supuesto de normalidad, normalidad de errores, diseño experimental, distribución normal, distribución uniforme

ABSTRACT

Theoretically linear models assume normal errors. This is not always true in real life, therefore the purpose of this study is to observe what happens when this rule is not followed. For example, when the data found is not normal but instead uniform, log-normal or exponential. For this reason, we want to simulate four models with three treatments each, with the purpose of extracting and analyzing the power of the test for the difference of means. For tests a level of significance of 5% is established.

KEY WORDS: Difference of means, experimental design, normal distribution, uniform distribution,normality

INTRODUCCIÓN

Cuando hablamos de modelos para el análisis de datos, generalmente se piensa en modelos cuyos errores son independientes e idénticamente distribuidos de la forma Normal (0, $\sigma 2$). Sin embargo, la variable respuesta no siempre resulta tener esta distribución.

La distribución normal, cuya relevancia en estadística se debe a que muchos fenómenos físicos, biológicos, psicológicos o sociológicos, pueden ser adecuadamente modelizados mediante ella. (Tauber & Sánchez, 2001). Los teoremas de límite en cálculo de probabilidades aseguran que la media de las muestras aleatorias tienen una distribución aproximadamente normal para muestras de suficiente tamaño, incluso en poblaciones no normales, y muchos métodos estadísticos requieren la condición de normalidad para su correcta aplicación. (Tauber & Sánchez, 2001).

El objetivo de la investigación es comparar las diferencias que ocurren en tres modelos al obtener la potencia de la prueba cuando la distribución de la variable respuesta es normal y tres más cuando no es normal. Esta comparación genera la incógnita de cual modelo es mejor, según la

distribución que tienen las observaciones de la variable respuesta y se busca un resultado claro para esta respuesta.

La definición de la potencia de la prueba se encuentra en diversos artículos orientados a áreas de investigación que varían mucho entre sí pero con el mismo concepto en común, que define que "la potencia estadística constituye un índice de la validez de nuestros resultados estadísticos" (Cohen, 1992; Bono & Arnau Gras, 1995). Con respecto a los efectos de un análisis estadístico "podemos afirmar que cuanto mayor sea la muestra, mayor será la potencia estadística,(manteniendo constante el tamaño del efecto y α) dado que el error aleatorio de medida es menor" (Lipsey, 1990; Cohen, 1988).

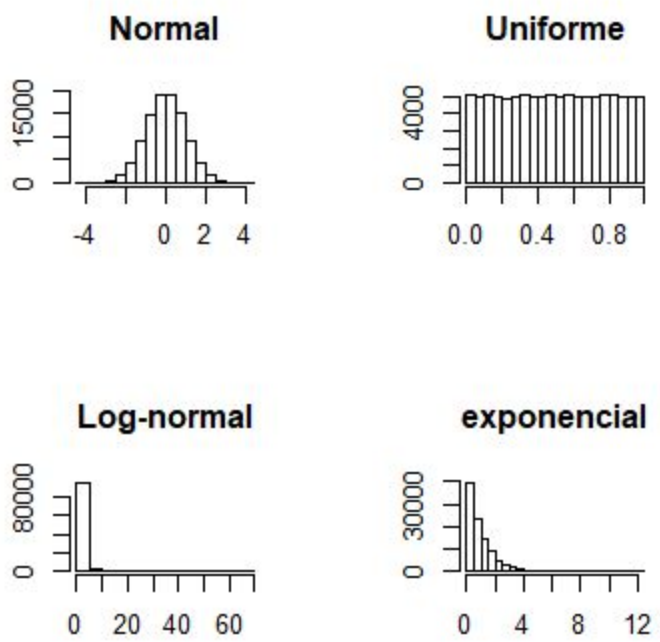
"El tamaño del efecto representa el grado en que la hipótesis nula es falsa. Cuando el tamaño del efecto es grande la potencia de la prueba aumenta" (Cohen, 1988, 1992). Por lo tanto para la investigación realizada se va a entender la potencia como "La sensibilidad de un diseño investigativo, reside en su capacidad de detectar diferencias o efectos allí donde los haya" o dicho de otra manera "La potencia puede ser concebida como la probabilidad de rechazar estadísticamente la hipótesis nula cuando ésta es falsa". Es decir, cuán probable es que los investigadores demuestran estadísticamente que su hipótesis inicial era correcta.

Y por último para adoptar el uso de la potencia de la prueba en todas la investigaciones que sea posible, tomar en cuenta que "una primera y extremadamente útil aplicación de la potencia estadística tiene que ver con la posibilidad de determinar, *a priori,* el tamaño muestral requerido para que la investigación tenga una potencia aceptable y "la potencia deseada va a ser siempre 0.8 o más " según Cohen(1992).

METODOLOGÍA

Se busca analizar cómo se comportan los modelos experimentales cuando no se logra cumplir el supuesto de normalidad. La simulación se realizó bajo el concepto de un diseño experimental, simulando distintas variables respuestas, una con una distribución Normal y la otras tres simulando datos no normales, para este caso una distribución uniforme simétrica y la log-normal y la exponencial que son asimétricas. Se cuenta con un solo factor de diseño por modelo, con tres tratamientos. La simulación busca analizar el impacto que produce la violación del supuesto en el cálculo de la potencia de la prueba para encontrar diferencias en las 3 medias antes mencionadas utilizando tamaños de muestra de 6,9,12,15 y 30 para cada comparación, esto con un nivel de significancia de 0.05 para todos los casos.

Figura N°1. Comportamiento simétrico y asimétrico de las distribuciones utilizadas.



En la Figura N°1, se observa cómo tanto la distribución normal y la distribución uniforme poseen una distribución simétrica , caso contrario se observa en la distribución log-normal y la distribución exponencial las cuales poseen una forma asimétrica agrupando gran porcentaje de los datos al lado izquierdo de la distribución .

Para realizar la programación de la simulación se utilizó RStudio versión 1.2.1335 y R versión 3.6.1(RStudio Team, 2019). Se utilizó la librería “car” (John Fox and Sanford Weisberg, 2011) y se requieren las siguientes funciones:

- rnorm
- runif
- rlnorm
- rexp

- lm
- anova
- mean
- plot

Primeramente se creó una función llamada “datos” para simular cada una de las variables respuestas de manera aleatoria junto con una variable factor de 3 niveles. Seguidamente se construyen los cuatro modelos lineales correspondientes. Esta función finalmente extrae el valor-p de cada modelo.

Se programó una segunda función llamada “pot”, para cada uno de los modelos anteriormente creados, con los parámetros (tamaño de muestra ,selección de las tres medias correspondientes a cada tratamiento,varianza correspondiente) con el objetivo de repetir 200 veces la función “datos” y de esta manera promediar los valores p correspondientes devolviendo la potencia de la prueba para cada modelo creado con las especificaciones de los parámetros antes mencionados.

Finalmente se creó un vector con el resultado de la función “pot” y posteriormente se grafica para el análisis correspondiente utilizando una varianza de 5 y medias 10,15,20 para los 4 modelos.

RESULTADOS

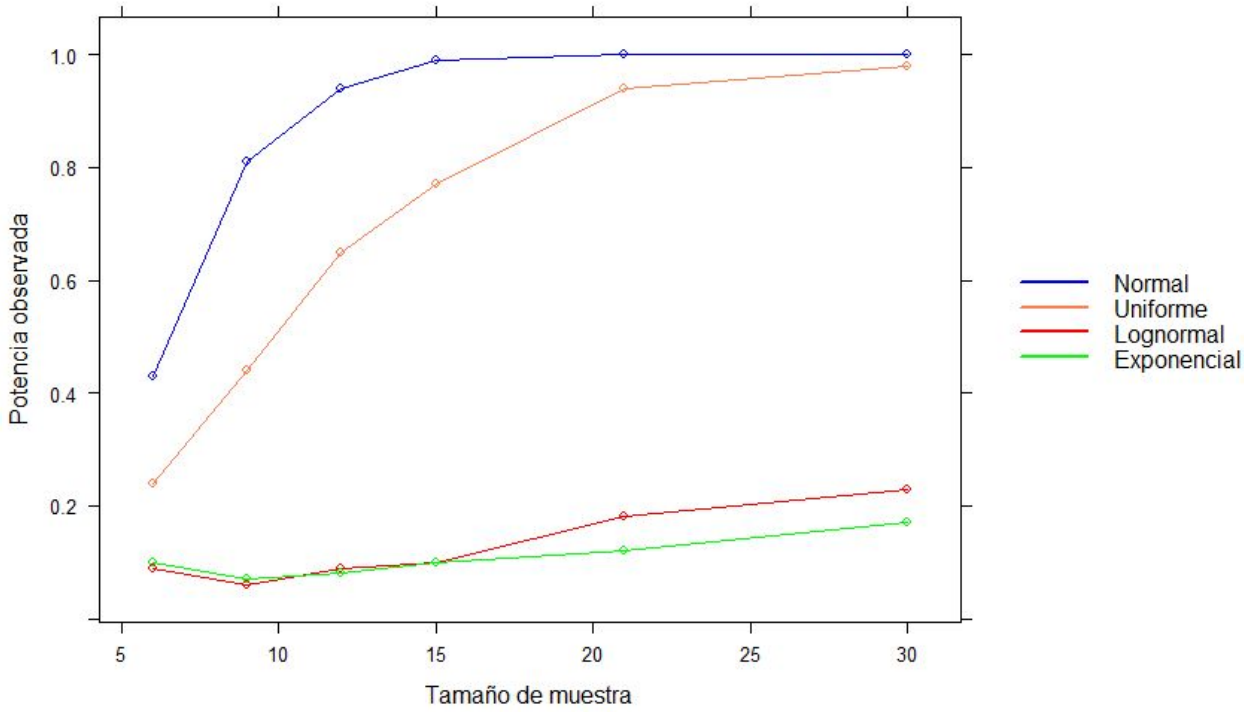
Los resultados obtenidos al correr la función que se programó para la simulación se muestran en la tabla de la Figura N°2

Figura N°2. Tabla de las potencias obtenidas según el tamaño de muestra, para cada distribución.

Tamaño de muestra	Normal	Uniforme	Log-normal	Exponencial
6	0.43	0.24	0.09	0.10
9	0.81	0.44	0.06	0.07
12	0.94	0.65	0.09	0.08
15	0.99	0.77	0.1	0.1
21	1.00	0.94	0.18	0.12
30	1.00	0.98	0.23	0.17

Básicamente las comparaciones deseadas de la potencia obtenida con cuatro modelos, cada uno con una variable respuesta de distinta distribución, al ir aumentando el tamaño de muestra y buscando el tamaño que proporcione una potencia deseada de 0.8 o más, con la intención de comparar las distribuciones tanto simétricas como asimétricas y si los resultados obtenidos tienen sentido con las definiciones utilizadas sobre la potencia de la prueba para así determinar si el uso de un modelo es más apropiado que algún otro.

Figura N°3. Gráfico de las potencias obtenidas por tamaño de muestra, agrupadas según la distribución.



En la Figura N° 3 se brinda un gráfico de los distintos valores de potencia obtenidos según el tamaño de muestra, agrupado por distribución, de manera que, se observa claramente que con las distribuciones simétricas utilizadas (normal y uniforme) se cumple que a mayor tamaño de muestra mayor potencia, aunque bien es cierto que la distribución normal alcanza la potencia deseada de 0.8 a menor muestra que la distribución uniforme, la cual si requiere mayor tamaño de muestra. Con respecto a las dos distribuciones asimétricas utilizadas en la simulación, se observa que la exponencial con una muestra de 6 proporciona mayor potencia que con una muestra de 9, aunque seguidamente a mayor tamaño de muestra si se obtiene una potencia mayor, aunque para obtener la potencia deseada o más si necesitara aún mucho más muestra que las dos distribuciones anteriores, y por último con la distribución log-normal, ocurre que con el menor tamaño de muestra se obtiene una mayor potencia que con una muestra más grande de 9 que se obtiene un menor tamaño de muestra, seguidamente vuelve a obtener una probabilidad mayor y se mantiene en constante crecimiento el valor de la potencia con mayor muestra, la distribución log-normal no cumple con la característica de una función monótona que solo crece o sólo decrece y que para la simulación con respecto a la potencia, se busca que siempre crezca para así cumplir con la definición utilizada de a mayor tamaño de muestra, mayor potencia.

CONCLUSIONES

Cuando la distribución es simétrica las diferencias que se encuentran entre ellas es por el tamaño de muestra necesario para lograr obtener la potencia deseada, mantienen entre sí una relación que cumple con el criterio de que a mayor tamaño de muestra, mayor potencia, aunque la distribución normal con los parámetros utilizados alcanza la potencia deseada con un tamaño de muestra de 9 y la

distribución uniforme con la misma relación de los parámetros de la normal, alcanza 0.8 con un tamaño de aproximadamente el doble pero con ambas se obtienen los resultados esperados.

En cuanto a las dos distribuciones asimétricas, rompen con la relación recién mencionada, ya que a mayor tamaño de muestra no necesariamente se obtiene una potencia mayor, con la distribución exponencial, con la muestra de 6 el valor de la potencia es mayor que con la muestra de 9, seguidamente de este tamaño de muestra, sí se cumple la relación y el hecho de que no cumpla estrictamente con la curva creciente tomando como guía la Figura N°3 se puede estar dando por la semilla utilizada para la simulación o la condición de utilizar tamaños de muestra que son relativamente pequeños, ya que se sabe que las distribuciones asimétricas generan problemas cuando la muestra es pequeña. Con la distribución log-normal, los resultados obtenidos se asemejan mucho al comportamiento que tiene la distribución exponencial, y el hecho de que la potencia es menor con el tamaño de muestra 9 es muy probablemente por la misma razón que en la exponencial.

El modelo que proporcionó los resultados más útiles es el que utiliza la variable respuesta con distribución normal como era de esperar, aunque también la distribución uniforme brinda resultados muy importantes, ya que a pesar de ser un modelo que no cumple con el supuesto de normalidad, cumple con los estándares esperados con el tema de la potencia, y aunque crece más lento que la normal, se logran los resultados meta. La literatura no indica que las distribuciones asimétricas rompen con el esquema planteado de la potencia de la prueba, pero los resultados obtenidos demuestran que dichas distribuciones no son útiles para lograr aumentar la potencia con un tamaño de muestra similar con el que las otras dos distribuciones simétricas logran alcanzar la potencia deseada.

BIBLIOGRAFÍA

Batanero, Carmen; Tauber, Liliana M; Sánchez, Victoria (2001). *Significado y comprensión de la distribución normal en un curso introductorio de análisis de datos* . Research Gate. URL: https://www.researchgate.net/profile/Carmen_Batanero/publication/282281515_Comprension_de_la_distribucion_normal_por_estudiantes_univesitarios/links/574ebfe808ae789584d7b24c.pdf

Cohen, J, (1988). *Statistical Power Analysis for the Behavioral Sciences.* (2nd ed.), New Jersey: Lawrence Erlbaum Associates.

Cohen, J. (1992). *Cosas que he aprendido (hasta ahora)*. Anales de Psicología, 8(1- 2), 3-18.

García, J. Fernando. Pascual, Juan. Dolores Frías,María . Dirk Van Krunckelsven* y Murgui, Sergio . *"Diseño y análisis de la potencia: n y los intervalos de confianza de las medias*." Universidad Católica de Valencia (2008) https://www.redalyc.org/html/727/72720464/

Quezada, Camilo. *Potencia Estadística, Sensibilidad y Tamaño del Efecto: ¿Un nuevo canon para la investigación?*. Pontificia Universidad Católica.

John Fox and Sanford Weisberg (2011). *An {R} Companion to Applied Regression,* Second Edition. Thousand Oaks CA: Sage . URL: http://socserv.socsci.mcmaster.ca/jfox/Books/Companion

RStudio Team (2019). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*, Vienna, Austria.URL https://www.R-project.org/.

ANEXOS

```
 datos_n<-function(n,m1,m2,m3,v){
  n <- as.integer(n)
  x = factor(rep(1:3, each=n/3))

  ynorm <- c(rnorm(n/3,m1,v), rnorm(n/3,m2,v), rnorm(n/3,m3,v))
  modn <-lm(ynorm~x)

  p <- anova(modn)[1,5]
  return(p)
}
datos_u<-function(n,m1,m2,m3,v){
  x = factor(rep(1:3, each=n/3))
  n <- as.integer(n)
 yunif <- c(runif(n/3,m1-(m1/2), m1+(m1/2)), runif(n/3,m2-(m2/2), m2+(m2/2)),
runif(n/3,m3-(m3/2), m3+(m3/2) ))
  modu <-lm(yunif~x)
  p <- anova(modu)[1,5]
  return(p)
}
datos_l<-function(n,m1,m2,m3,v){
  x = factor(rep(1:3, each=n/3))
  n <- as.integer(n)
 ylnorm <- c(rlnorm(n/3,m1,v),rlnorm(n/3,m2,v),rlnorm(n/3,m2,v))
  modl <-lm(ylnorm~x)
  p <- anova(modl)[1,5]
  return(p)
}
datos_e<-function(n,m1,m2,m3,v){
  x = factor(rep(1:3, each=n/3))
  n <- as.integer(n)
 yexp <- c(rexp(n/3,1/m1),rexp(n/3,1/m2),rexp(n/3,1/m3))
  mode <-lm(yexp~x)
  p <- anova(mode)[1,5]
  return(p)
}
pot_n=function(n,m1,m2,m3,v,M){
  prob=c()
  for (j in 1:M) {
          prob <- c(prob, datos_n(n,m1,m2,m3,v) )
  }
  pow=mean(prob<0.05)
```

```
  return(pow)
}
pot_u=function(n,m1,m2,m3,v,M){
  prob=c()
  for (j in 1:M) {
          prob <- c(prob, datos_u(n,m1,m2,m3,v) )
  }
  pow=mean(prob<0.05)
  return(pow)
}
pot_l=function(n,m1,m2,m3,v,M){
  prob=c()
  for (j in 1:M) {
          prob <- c(prob, datos_l(n,m1,m2,m3,v) )
  }
  pow=mean(prob<0.05)
  return(pow)
}
pot_e=function(n,m1,m2,m3,v,M){
  prob=c()
  for (j in 1:M) {
          prob <- c(prob, datos_e(n,m1,m2,m3,v) )
  }
  pow=mean(prob<0.05)
  return(pow)
}
pot(n, media1, media2, media3, varianza, M)
pot_n(20,0,5,10,5,200)
## [1] 0.805
pot_u(20,0,5,10,5,200) #no usa la varianza
## [1] 1
pot_l(20,0,5,10,5,200)
## [1] 0.015
pot_e(20,0,5,10,5,200) #no usa la varianza
## [1] 0.67
n <-rep(c(6,9,12,15,21,30),each=4)
pot1<-c(0.43,0.24,0.06,0.10,0.81,0.44,0.02,0.07,0.94,0.65,0.03,0.08,0.99,0.77,0.01,0.1,1.00,0.9
4,0.05,0.12,1.00,0.98,0.02,0.17)
dist <- rep(c(1,2,3,4),6)
base=data.frame(cbind(n,pot1,dist))
base$dist=factor(base$dist)
levels(base$dist)=c("Normal","Uniforme","Lognormal","Exponencial")
attach(base)
```

```
## The following objects are masked _by_ .GlobalEnv:
##
##       dist, n, pot1
str(base)
## 'data.frame':          24 obs. of  3 variables:
##  $ n   : num  6 6 6 6 9 9 9 9 12 12 ...
##  $ pot1: num  0.43 0.24 0.06 0.1 0.81 0.44 0.02 0.07 0.94 0.65 ...
##  $ dist: Factor w/ 4 levels "Normal","Uniforme",..: 1 2 3 4 1 2 3 4 1 2 ...
library(lattice)

xyplot( pot1~n,
          groups = dist,
          type="b",
          points = FALSE,
          lines = TRUE , col=c("blue","coral","red","green"),
          xlab= "Tamaño de muestra",
          ylab = "Potencia observada",

          key=list(space="right",
          lines=list(col=c("blue","coral","red","green"), lty=c(1,1,1,1), lwd=2),
          text=list(c("Normal","Uniforme","Lognormal","Exponencial")))
          )
```