Análisis de Regresión (2021-3)

LOS LIBERTADORES
FUNDACIÓN UNIVERSITARIA

Especialización en Estadística Aplicada

Prof. Sébastien Lozano Forero (slozanof@libertadores.edu.co)

Estadística Inferencial

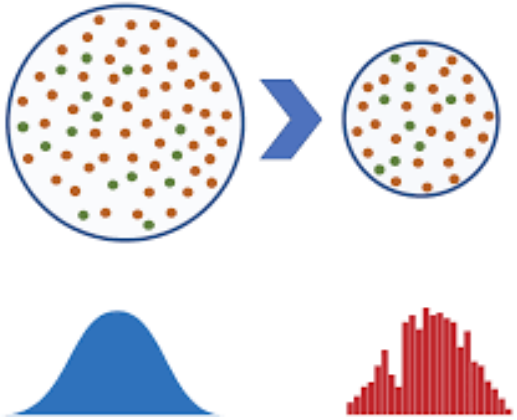
Tabla de contenidos

Presentación

Población: conjunto de individuos, objetos o fenómenos de los cuales se desea estudiar una o varias características.

muestra: subconjunto de casos o individuos de una población. En diversas aplicaciones, interesa que una muestra sea representativa, y para ello debe escogerse una técnica de muestra adecuada que produzca una muestra aleatoria adecuada. (Esto se estudiará en el espacio académico de *muestreo estadístico* del programa).

La estadística inferencial permite obtener conclusiones a partir de una **muestra**, pero que son aplicables a una **población**. Es necesario tener presente que la diferencia principal entre la matemática y la estadística yace en el manejo del error (en matemáticas se asume de forma determinística, mientras que en estadística se asume de forma estocástica), de esta manera, el espíritu central de la Estadística Inferencial yace en *tomar decisiones en escenarios de incerteza*.



Estimación Puntual

"Dios es el único que conoce los parámetros poblacionales, pero no nos dirá cuáles son".

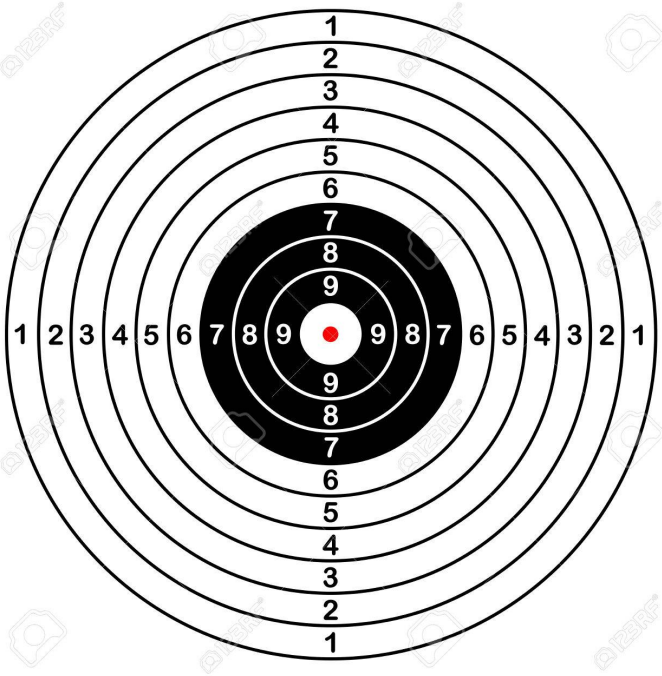
Los parámetros poblacionales suponen, en una mayoría considerable de casos, el objeto de deseo máximo en estadística. Sin embargo, no es posible obtenerlos. Por tanto, una herramienta importante para **estimar** parámetros es el uso de la información disponible en la muestra. De esta manera, el problema se traduce en usar la información disponible de la mejor manera posible. En general la situación se puede entender como:

$$\begin{cases} X_1, X_2, \cdots, X_n \sim f(\theta), \\ \theta: \text{Parámetro (escalar). } \theta \in \mathrm{R}, \text{ (puede ser vector)}, \\ \hat{\theta}: \text{Estimador (Variable Aleatoria)}, g(X_1, \cdots, X_n) \text{ con } g: \mathrm{R}^n \to \mathrm{R}. \\ f_{\hat{\theta}}(x): \text{Función de densidad de probabilidad de } \hat{\theta} \end{cases}$$

De esta manera, se tienen varias definiciones y características

- **Sesgo de un estimador.** $\mathrm{B}(\hat{\theta}) = E(\hat{\theta}) - \theta$
- **Error Cuadrático medio.** $\mathrm{EQM}(\hat{\theta}) = \mathrm{Var}(\hat{\theta}) + \mathrm{B}^2(\hat{\theta})$
- **Función de verosimilitud** $L(\theta) = \Pi_{i=1}^{n} f_{X_i}(x_i)$
- **Función de log-verosimilitud** $\ell(\theta) = \ln(L(\theta))$
- **Estimador consistente.** $\lim_{n \to \infty} \mathrm{Var}(\hat{\theta}) = 0$

- **Estimador eficiente.** $Var(\hat{\theta}) = \left\{ E\left(\frac{\partial}{\partial\theta} \ell(\theta) \right)^2 \right\}^{-1}$ (Desigualdad de Cramer-Rao)
- **Estimador Suficiente.** $P(X_1, \cdots, X_n \,|\, \theta, \hat{\theta}) = P(X_1, \cdots, X_n \,|\, \hat{\theta})$ (Criterio de Factorización de Fisher-Neymman)



Ahora, todas las definiciones anteriores, en un sentido práctico, nos dicen que la información se está utilizando de la mejor manera posible. Por tanto, podemos confiar en las estimativas númericas (tomar la fórmula del estimador y aplicarla con un conjunto determinado de datos) del estimador.

Por ejemplo, la **media muestral** (sean $X_1, \cdots, X_n$ una muestra aleatoria, $\bar{X} = n^{-1} \sum_{i=1}^{n} X_i$) es el mejor estimador posible para la **media poblacional** (μ) y la **varianza muestral** $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$ es el mejor estimador posible para la **varianza poblacional** (σ^2). Veamos:

In [ ]:
```
set.seed(1234) #semilla 1
# generación de datos aleatorios
x <- rnorm(200, mean=100, sd=81) # 200 datos aleatorios de una dist normal con media 100 y desviación 81.
# # parámetro poblacional mu=100 y sigma2=81
mean(x) # estimación de la media
sd(x) # estimación de la desviación
# # estimadores Xbarra=95.3214 y sigma2=82.6739
```

In [ ]:
```
set.seed(4321) #semilla 2
# generación de datos aleatorios
y <- rnorm(200, mean=100, sd=81) # 200 datos aleatorios de una dist normal con media 100 y desviación 81.
# parámetro poblacional mu=100 y sigma2=81
mean(y) # estimación de la media
sd(y) # estimación de la desviación
# estimadores Xbarra=104.2923 y sigma2=75.0056
```

Misma ecuación, pero resultados diferentes. Así se ven las dos muestras (que vienen de la misma población y corresponden a los mismos

parámetros poblacionales)

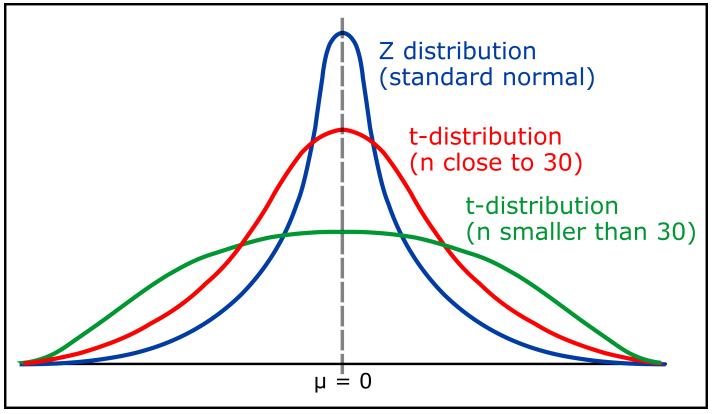
In [ ]:
```
#librerías
# install.packages("ggplot2")
library(ggplot2)
library(repr)
options(repr.plot.width=6, repr.plot.height=6)
```

In [ ]:
```
#data frame para los datos
data <- data.frame(tipo = c( rep("x", 200), rep("y", 200) ),valor = c(x,y))
# visualización
ggplot(data, aes(valor, fill = tipo)) + geom_density(alpha = 0.2)
```

Distribuciones Muestrales

"La distribución Normal es la distribución favorita de diosito".

Al momento de tomar una muestra, asumimos que existirá un error asociado al hecho de no estar utilizando toda la información (población) sino una parte de ella. Adicionalmente, la intención principal es poder establecer hechos sobre estadísticos destacados. Por tanto, es de interés poder usar la información de las variables que componen la muestra en favor de conocer algún supuesto distribucional de estimadores populares. Estas distribuciones serán fundamentales para poder incoporar a nuestra batería de herramientas los intervalos de confianza y las pruebas de hipótesis.



Ejemplo de la importancia de tales distribuciones, es el **Teorema del Límite Central (TLC)** que establece:

$$X_1, X_2, \cdots, X_n \sim f(\mu, \sigma^2), \text{ (f es cualquier distribución), entonces } \bar{X} \sim N(\mu, \sigma^2/n)$$

Nuestro interés será utilizar diferentes distribuciones muestrales para poder establecer relaciones entre probabilidades y cuantiles.

Sea $X_1, \cdots, X_n \sim N(\mu, \sigma^2)$, las siguientes son resultados o distribuciones muestrales asociadas a la distribución normal.

- **Normal.** $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$

Sea $Z \sim N(0,1)$. Así, $P(Z > z_\alpha) = \alpha$, donde z_α es cuantil y α es probabilidad

In [ ]:
```
normal <- rnorm(1000000) #1000 datos de una distribución normal estándar
hist(normal, prob=TRUE, col="grey", main ="Histograma de la dist normal", ylab="Densidad")
lines(density(normal), col="blue", lwd=2)
lines(density(normal, adjust=2), lty="dotted", col="darkgreen", lwd=2)
```

- **t-student.** $T = \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{(n-1)}$

  Sea $T \sim t_{(n-1)}$. Así, $P(T > t_{(n),\alpha}) = \alpha$, donde $T_{(n),\alpha}$ es cuantil y α es probabilidad

In [ ]:
```
T <- rt(10000, df=25) #1000 datos de una distribución t
hist(T, prob=TRUE, col="grey", main ="Histograma de la dist t-student", ylab="Densidad")
lines(density(T), col="blue", lwd=2)
lines(density(T, adjust=2), lty="dotted", col="darkgreen", lwd=2)
```

- **chi-cuadrado** $\chi = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{(n-1)}$

  Sea $\chi \sim \chi^2_{(n-1)}$. Así, $P(\chi > \chi^2_{(n),\alpha}) = \alpha$, donde $\chi^2_{(n),\alpha}$ es cuantil y α es probabilidad

In [ ]:
```
chi <- rchisq(100000, df=25) #10000 datos de una distribución chi
hist(chi, prob=TRUE, col="grey", main ="Histograma de la dist Chi cuadrado", ylab="Densidad") # prob=TRUE for probabi
lines(density(chi), col="blue", lwd=2) # add a density estimate with defaults
lines(density(chi, adjust=2), lty="dotted", col="darkgreen", lwd=2)
```

Sean $X_1, \cdots, X_n \sim N(\mu_X, \sigma^2_X)$ y $Y_1, \cdots, Y_m \sim N(\mu_Y, \sigma^2_Y)$ la siguiente es una distribución muestral asociadas a la distribución normal.

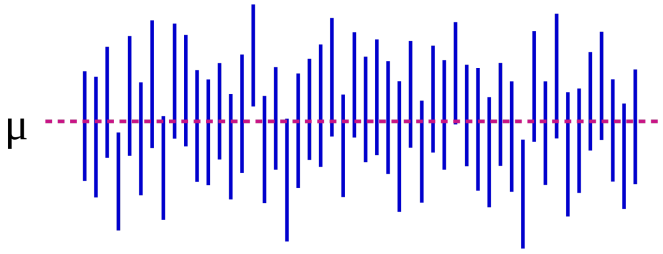
- **F** $F = \left(\frac{S_X}{S_Y}\right)^2 \left(\frac{\sigma_Y}{\sigma_X}\right) \sim F_{(n-1,m-1)}$

  Sea $F \sim F_{(n-1,m-1)}$. Así, $P(F > F_{(n-1,m-1),\alpha}) = \alpha$, donde $\chi^2 F_{(n-1,m-1),\alpha}$ es cuantil y α es probabilidad.

In [ ]:
```
F <- rf(1000000, df1=25, df2=20) #10000 datos de una distribución F
hist(F, prob=TRUE, col="grey", main ="Histograma de la dist F", ylab="Densidad") # prob=TRUE for probabilities not co
lines(density(F), col="blue", lwd=2) # add a density estimate with defaults
lines(density(F, adjust=2), lty="dotted", col="darkgreen", lwd=2)
```

Intervalos de Confianza

Los intervalos de confianza son la "evolución" natural de la estimación puntual al tratarse de una metodología que no intenta gastar "su única bala" estimando un único parámetro, sino que intenta barrer sobre un intervalo numérico para, con una probabilidad razonablemente alta ($1-\alpha$, donde α es llamada de *significancia*), el intervalo puede contener al parámetro. Esto es,



$$\begin{cases} X_1, X_2, \cdots, X_n \sim f(\theta), \\ \theta: \text{Parámetro (escalar). } \theta \in \mathrm{R}, \\ \hat{\theta}_U, \hat{\theta}_L: \text{Estimadores (en este contexto, límites inferior y superior, respectivamente).} \\ \alpha \in (0,1): \text{Significancia} \end{cases}$$

Decimos que $\left(\hat{\theta}_U, \hat{\theta}_L\right)$ es un Intervalo de Confianza (IC) para θ, con confianza del $100(1-\alpha)\%$, si $P\left(\theta \text{ esté contenido en } (\hat{\theta}_U; \hat{\theta}_L)\right) = 1-\alpha$.

En general, las distribuciones muestrales son herramientas con las que se dota de aleatoriedad a los Intervalos de confianza. Así, regularmente en Estadística Inferencial se estudia la construcción de Intervalos de confianza clásicos, a partir de las distribuciones muestrales. Algunos intervalos particulares:

- Un IC para μ con σ^2 conocido. $(X_1, X_2, \cdots, X_n \sim N(\mu, \sigma^2))$, con el $100(1-\alpha)\%$ de confianza, está dado por $\left(X \pm \frac{Z_{\alpha/2}\sigma}{\sqrt{n}}\right)$. Es decir,

$$P\left(\left(X \pm \frac{Z_{\alpha/2}\sigma}{\sqrt{n}}\right) \text{contenga a } \mu\right) = 1-\alpha$$

In [ ]:

```
#Validación por ciclos Monte Carlo. (ver https://en.wikipedia.org/wiki/Monte_Carlo_method)
cont <- 0 #contador
mu <- 100 #parámetro en la simulación
alpha <- 0.05 #nivel de significancia
sigma <- 10 #desviación estándar
n <- 100 #tamaño de muestra
for(i in 1:10000){#10000 réplicas
muestra <- rnorm(n, mean=mu, sd=sigma) #mu=100 (desconocido) y sd= 10 (conocido)
Xbarra <- mean(muestra)
s <- sd(muestra)
LI <- Xbarra-qt(alpha/2, df= n-1, lower.tail=F)*s/sqrt(n)
LS <- Xbarra+qt(alpha/2, df= n-1, lower.tail=F)*s/sqrt(n)
if(LI <= mu & mu <= LS){
    cont <- cont+1
}
}
print(paste("El intervalo contuvo al parámetro el", cont/100, "% de las réplicas "))
```

- Un IC para μ con σ^2 desconocido. $(X_1, X_2, \cdots, X_n \sim N(\mu, \sigma^2))$, con el $100(1-\alpha)\%$ de confianza, está dado por $\left(X \pm \frac{t_{(n-1),\alpha/2}S}{\sqrt{n}}\right)$

- Un IC para σ^2 con μ conocido. $(X_1, X_2, \cdots, X_n \sim N(\mu, \sigma^2))$, con el $100(1-\alpha)\%$ de confianza, está dado por $\left(\frac{(n-1)S^2}{\chi^2_{(n),\alpha/2}}; \frac{(n-1)S^2}{\chi^2_{(n),1-\alpha/2}}\right)$

- Un IC para σ^2 con μ desconocido. $(X_1, X_2, \cdots, X_n \sim N(\mu, \sigma^2))$, con el $100(1-\alpha)\%$ de confianza, está dado por $\left(\frac{(n-1)S^2}{\chi^2_{(n-1),\alpha/2}}; \frac{(n-1)S^2}{\chi^2_{(n-1),1-\alpha/2}}\right)$

- Un IC para p. (El estimador es $\hat{p} = X/n$, donde $X \sim Bn(n, p)$), con el $100(1-\alpha)\%$ de confianza, está dado por $\left(\hat{p} \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$

Pruebas de Hipótesis

"Todos somos inocentes hasta que se demuestre lo contrario".

Método Científico

i) **Plantear teorías**. La hipótesis es la explicación que se da a partir de las observaciones realizadas. De este modo, se presenta como una posible teoría. Sin embargo, habrá que tener en cuenta que una hipótesis siempre será una posibilidad, pero que será necesario reforzar mediante nuevos estudios, para lo que será necesario llevar a cabo una serie de experimentos.

ii) **Recopilar evidencias**. Este paso es posterior a la hipótesis y su función principal será darle validez mediante experimentos que sirvan para demostrar la veracidad de la hipótesis planteada. En el caso de que los experimentos lleven a negar la hipótesis, será necesario descartarla y formular una nueva hipótesis que responda de forma satisfactoria a las observaciones llevadas a cabo durante la experimentación y la observación.

iii) **Tomar una decisión**. Una vez que la experimentación haya servido para demostrar que la hipótesis planteada tiene sentido, se elaborará una teoría. La teoría será el resultado de aquellas hipótesis que tengan una probabilidad mayor de ser confirmadas como ciertas.

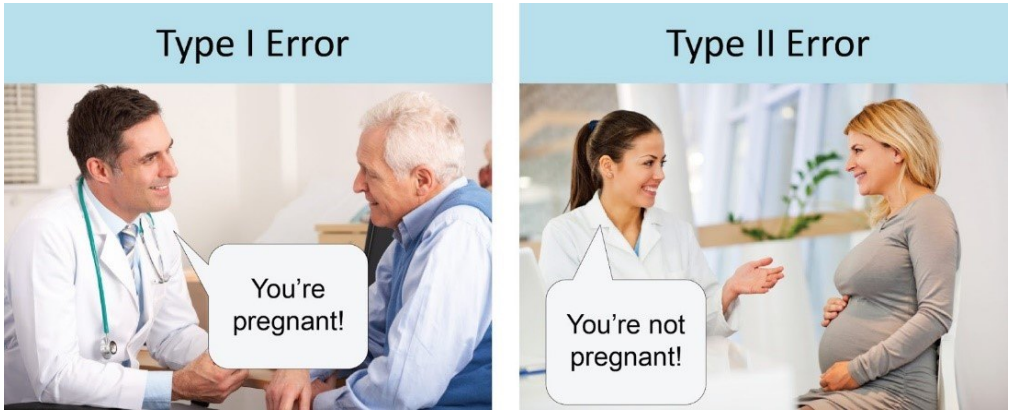
Elementos de una Prueba de Hipótesis

1. Hipótesis (suponen una partición de la realidad)

$$\begin{cases} H_0: \text{Hipótesis Nula} \\ H_1: \text{Hipótesis Alternativa} \end{cases}$$

2. Dada una muestra, se obtiene el estadístico $\tau \overset{H_0}{\sim}$ Distribución nula y $\tau \overset{H_1}{\sim}$ Distribución Alternativa. Como asumimos H_0 como verdadera, nos quedaremos también con la distribución nula hasta que se demuestre lo contrario.
3. Regla de decisión: p-valor (probabilidad) o valor crítico (cuantil)

Ahora, en últimas, las pruebas de hipótesis son usadas para tomar decisiones.



De esta manera, sean:

- $\alpha = P(\text{Error Tipo 1}) = P(\text{Rechazar } H_0 \,|\, H_0 \text{ es Verdadera})$
- $\beta = P(\text{Error Tipo 2}) = P(\text{No rechazar } H_0 \,|\, H_0 \text{ es falsa})$

α es llamada de **significancia estadística** y $1 - \beta$ es llamado de **Poder de la prueba**.

En general, para aplicar pruebas de hipótesis, basta tener presente los tres pasos fundamentales. Todas las pruebas de hipótesis se pueden entender de esa manera. Algunas pruebas para una población:

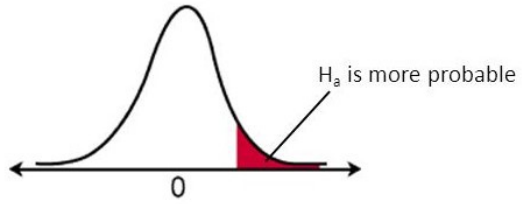
- Una prueba para μ con σ^2 desconocido. $(X_1, X_2, \cdots, X_n \sim N(\mu, \sigma^2))$, con $100\alpha\%$ de significancia, está dado por

i. Planteamiento de Hipótesis

$$\begin{cases} H_0 : \mu = \mu_0 (\mu_0 \text{ es conocido}) \\ H_1 : \begin{cases} \mu > \mu_0 (\text{ Prueba a cola derecha [Caso 1]}) \\ \mu < \mu_0 (\text{ Prueba a cola izquierda [Caso 2]}) \\ \mu \neq \mu_0 (\text{ Prueba bicaudal [Caso 3]}) \end{cases} \end{cases}$$

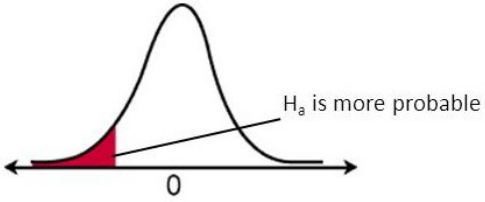
ii. Estadístico de prueba. $Z_{Est} = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} \overset{H_0}{\sim} N(0, 1)$

iii. Regla de Rechazo. Hay evidencia significativa para el rechazo de H_0 en favor de H_1 con $100\alpha\%$ si $p -$ valor $< \alpha$. Suponga que $Z \sim N(0, 1)$, entonces $p -$ valor $= P(Z > |Z_{Est}|)$ para los casos 1 y 2; $p -$ valor $= 2P(Z > |Z_{Est}|)$ para el caso 3.



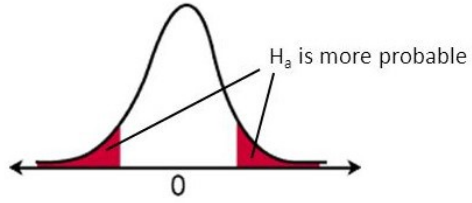
Right-tail test

$H_a: \mu > \text{value}$



Left-tail test

$H_a: \mu < \text{value}$



Two-tail test

$H_a: \mu \neq \text{value}$

Evaluemos la prueba de Hipótesis $H_0 : \mu = 25$ vs $H_1 : \mu \neq 25$ con $\bar{X} = 22$, $\sigma = 2$, $n = 40$ y $\alpha = 0.05$

In [ ]:
```
#Validación por ciclos Monte Carlo.
```

```r
cont <- 0 #contador
mu0 <- 25 #parámetro en la simulación (H0 es verdadera)
alpha <- 0.1 #nivel de significancia
sigma <- 2 #desviación estándar
n <- 40 #tamaño de muestra
for(i in 1:10000){#10000 réplicas
muestra <- rnorm(n, mean=mu0, sd=sigma) #mu=100 (desconocido) y sd= 10 (conocido)
Xbarra <- mean(muestra)
Zest <- (Xbarra - mu0)/(sigma/sqrt(n))
p.val <- 2*pnorm(abs(Zest), lower.tail=F)
if(p.val<alpha){
    cont <- cont+1
}
}
print(paste("El ciclo rechazó el", cont/100, "% de las pruebas "))
```

Pruebas de hipótesis no paramétricas

$$\begin{cases} H_0: \text{Los datos son normalmente distribuidos} \\ H_1: \text{Los datos no son normalmente distribuidos (no } H_0) \end{cases}$$

In [ ]:
```r
library(tseries)
jarque.bera.test(rchisq(1000, df=20))
```

In [ ]:
```r

```