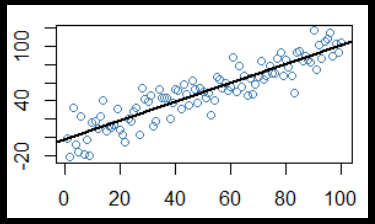
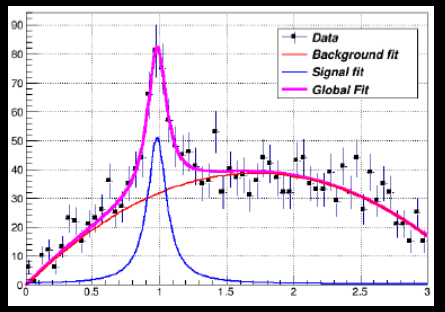
Temas avanzados en física computacional
Análisis de datos





Semestre

2016-I

Clase-5

José Bazo

jbazo@pucp.edu.pe

Contenidos del curso

- ✓ Introducción al análisis de datos y data science
- ✓ Lenguaje de programación R
- ✓ ROOT Data Analysis Framework
- ✓ Manipulación y visualización de datos

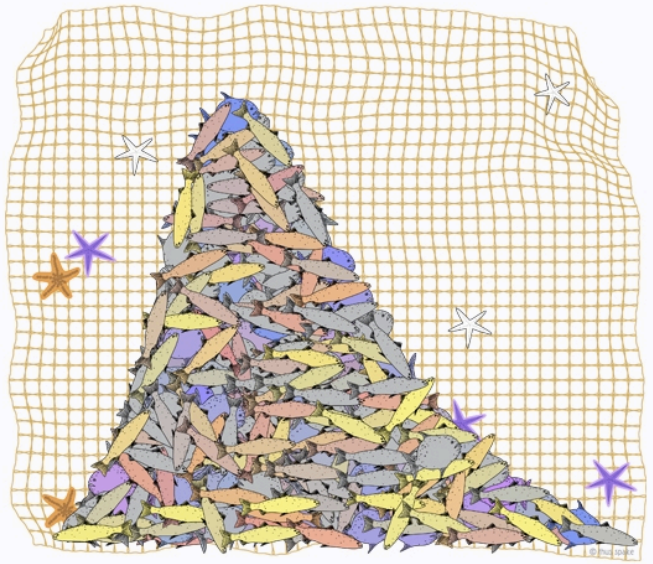
5. Modelamiento estadístico
6. Machine Learning
7. TMVA (Toolkit for Multivariate Data Analysis)

4. Modelamiento estadístico

Bibliografía

J. Canny. Introduction to Data Science . UC Berkeley

J. Akey. Introduction to Statistical Genomics. U Washington



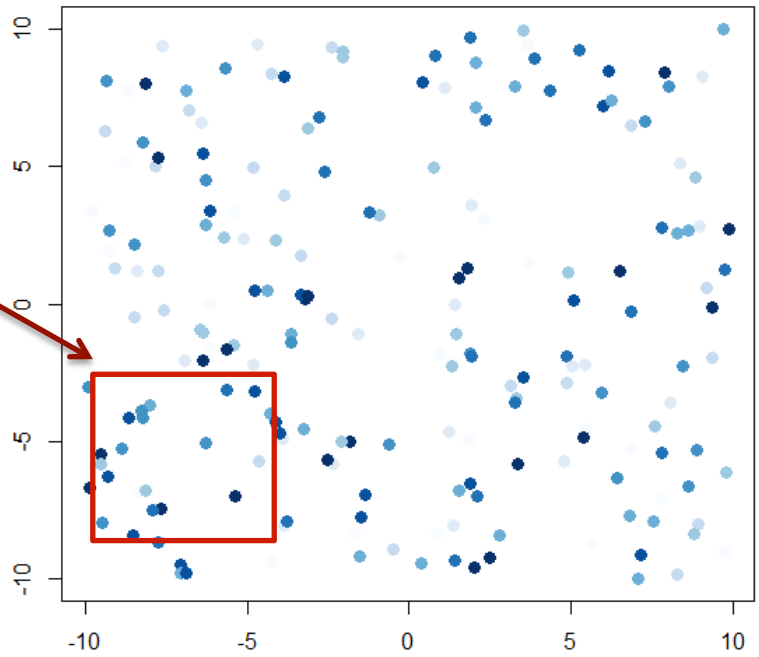
Medidas

Propiedades básicas:
Mínimo, máximo, promedio, mediana, moda, desviación estándar

Relaciones entre parámetros: gráfica de dispersión, regresiones, correlaciones

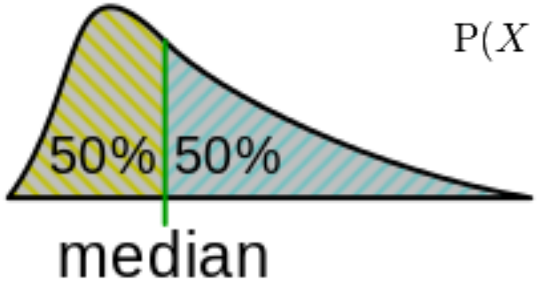
Juego de datos (**estadística**):
subconjunto de población mayor

-varían entre conjuntos (ruido, varianza)
-pueden ser sistemáticamente diferentes (bias)

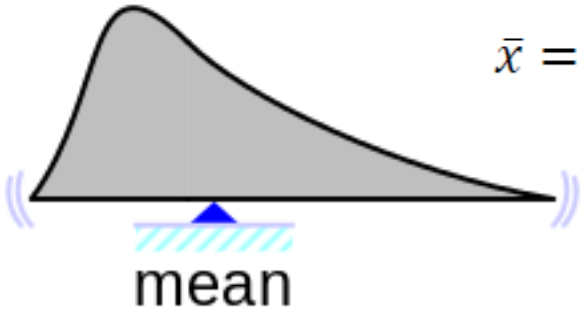


Media, Mediana, Moda

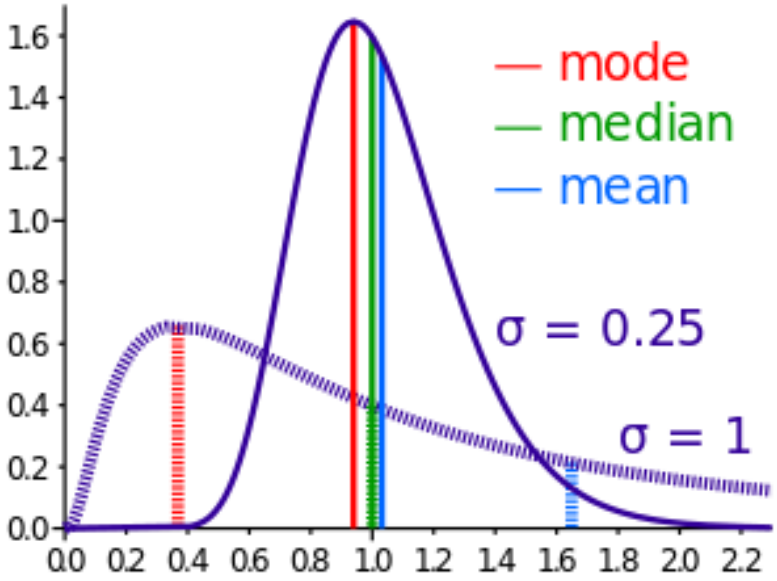
valor más común



$$\mathrm{P}(X \leq m) \geq \frac{1}{2} \text{ and } \mathrm{P}(X \geq m) \geq \frac{1}{2}$$



$$\bar{x} = \Sigma_{i=1}^{n} x_i / n$$



Varianza

Varianza poblacional:

$$\sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

Medida del ancho de la distribución

Desviación Estándar: σ

RMS (Root Mean Square)

$$x_{\mathrm{rms}}^2 = \bar{x}^2 + \sigma_x^2 = \overline{x^2}.$$

Varianza muestral:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2$$

Para variable aleatoria:

$$\sigma_X^2 = \mathrm{E}[(X - \mu)^2] \qquad \mu = \mathrm{E}[X],$$

$$\mathbb{E}[X] = \sum_{i=1}^{n} x_i p(x_i)$$

esperanza matemática, valor medio de un fenómeno aleatorio.

Covarianza

Covarianza entre dos variables aleatorias distribuidas conjuntamente (x vs y):

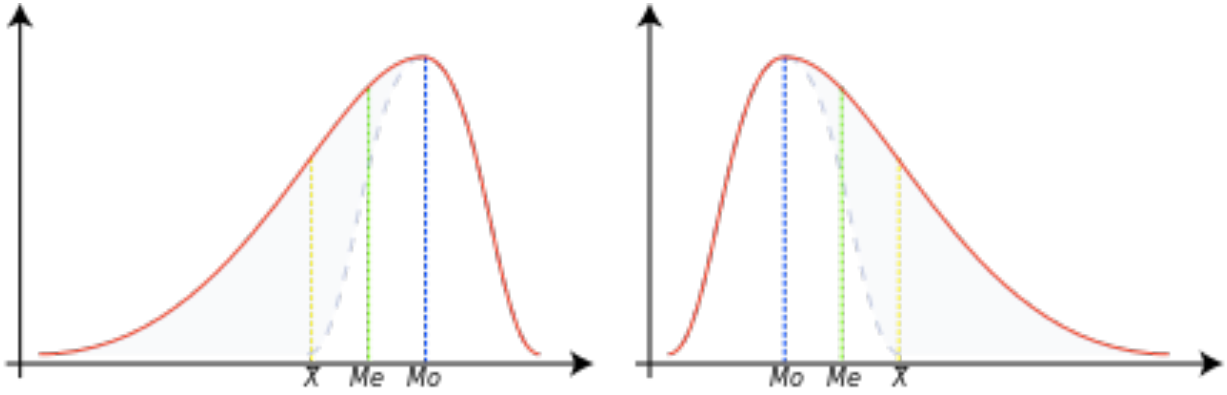
$$\operatorname{cov}(X,Y) = \operatorname{E}\left[(X - \operatorname{E}[X])(Y - \operatorname{E}[Y])\right]$$

En forma discreta:

$$\operatorname{cov}(X,Y) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{1}{2}(x_i - x_j)\cdot(y_i - y_j) = \frac{1}{n^2}\sum_{i}\sum_{j>i}(x_i - x_j)\cdot(y_i - y_j)$$

$$\operatorname{cov}(X,X) = \operatorname{Var}(X) \equiv \sigma^2(X).$$

Asimetría



Asimetría estadística (**skewness**): medida de asimetría de distribución de probabilidad

Coeficiente de asimetría de Pearson

$$A_p = \frac{\mu - moda}{\sigma}$$

válido para distribuciones uniformes, unimodales y moderadamente asimétricas

Coeficiente de asimetría de Fisher

$$\gamma_1 = \frac{\mu_3}{\sigma^3}: \qquad \gamma_1 > 0, \quad \text{a. positiva} \qquad \gamma_1 < 0, \quad \text{a. negativa}$$

momento central:

$$\mu_k = \mathrm{E}\left[(X - \mathrm{E}[X])^k\right] = \int_{-\infty}^{+\infty} (x-\mu)^k f(x)\,dx.$$

Distribución normal

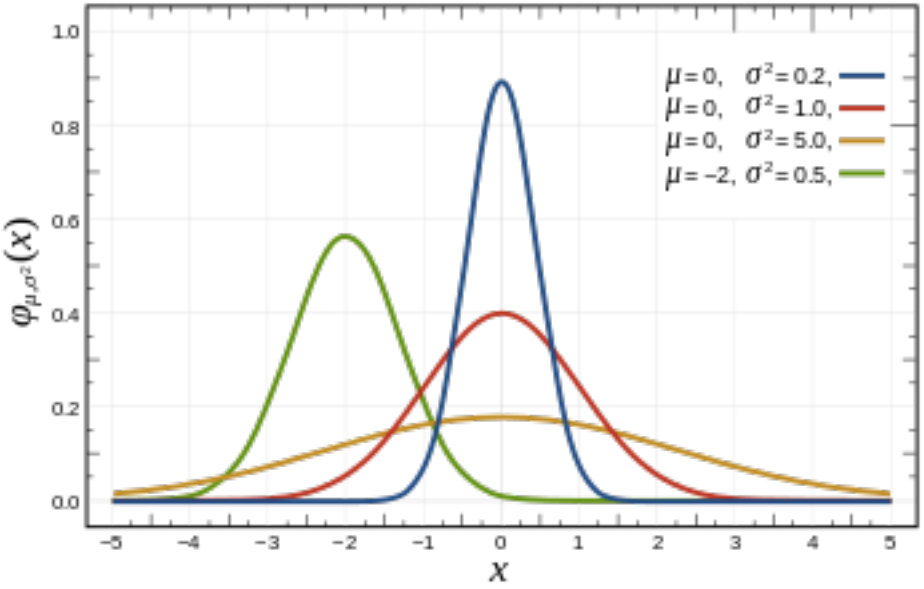
Normal o Gaussiana

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

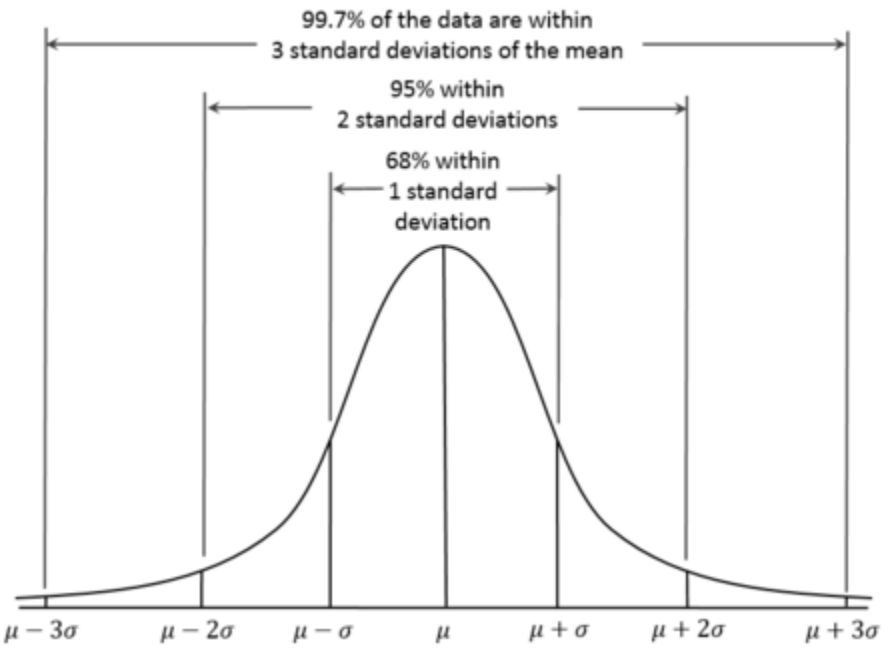
Media = Mediana = Moda = μ

Desviación Estándar = σ

Varianza = σ^2



Distribución normal



Intervalos de tolerancia

$$\Pr(\mu - \sigma \leq x \leq \mu + \sigma) \approx 0.6827$$
$$\Pr(\mu - 2\sigma \leq x \leq \mu + 2\sigma) \approx 0.9545$$
$$\Pr(\mu - 3\sigma \leq x \leq \mu + 3\sigma) \approx 0.9973$$

5σ ~ 0.9999994

En ciencias sociales 2σ podrían ser significativo como nivel de confianza.
En física de partículas:
- **evidencia**: 3σ
- **descubrimiento**: 5σ

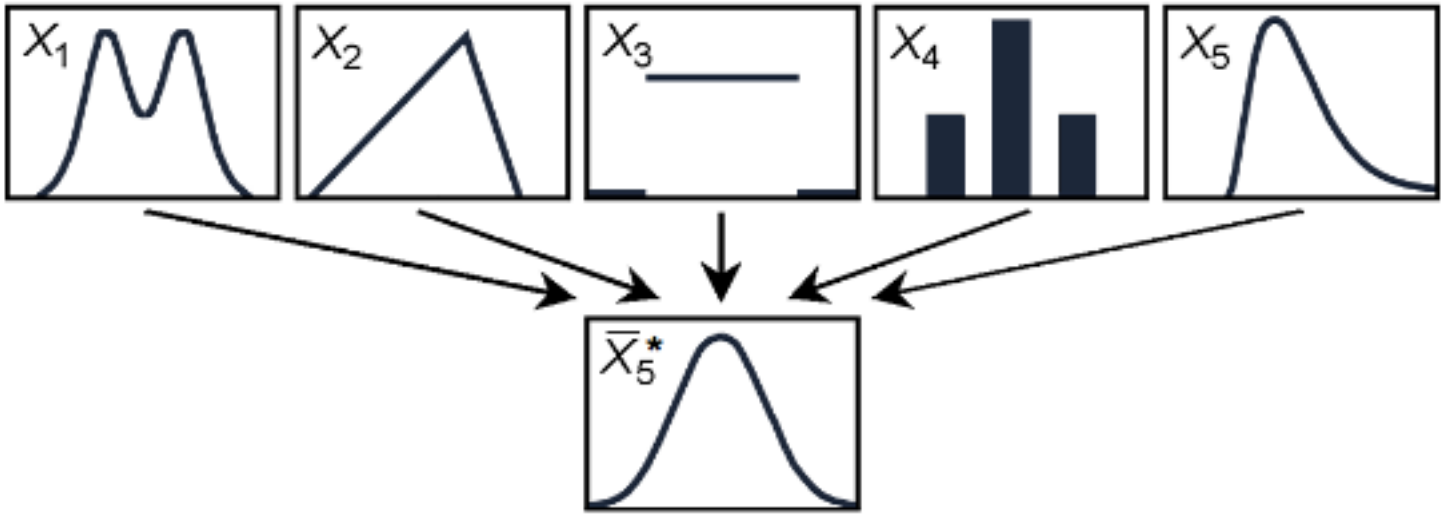
Teorema del límite central

Dado un conjunto de n variables aleatorias (X_i) e independientes de una distribución con media μ y varianza $\sigma^2 \neq 0$, luego cuando n -> ∞, la media aritmética

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

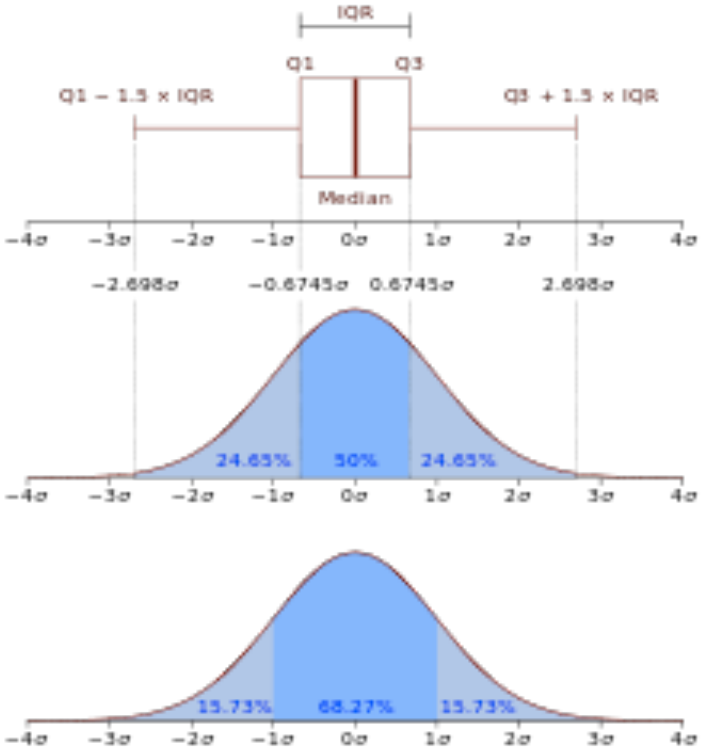
tiende a una distribución normal con $\mu_{\bar{X}} = \mu$ y $\sigma^2_{\bar{X}} = \frac{\sigma^2}{n}$.

Variables independientes aleatorias, con media 0 y σ=1



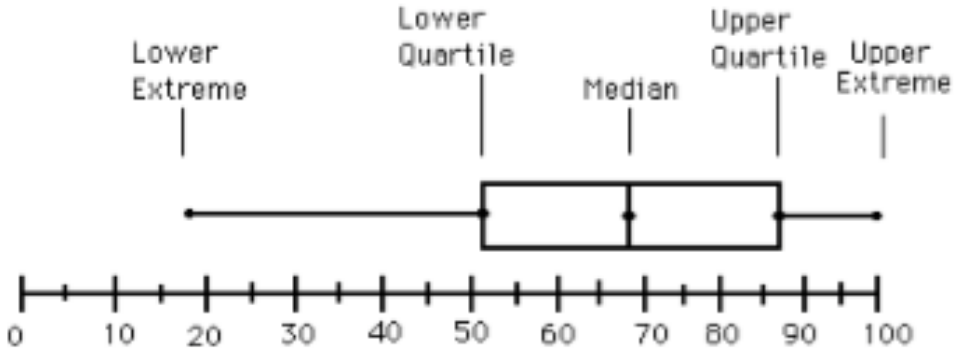
Corrigiendo Distribuciones

Muchas veces se asumen una distribución normal, pero a veces no es correcto



*IQR: interquartile range

Si la distribución es asimétrica no es normal



box-and-whisker plot

- Si x tiene una distribución log-normal, entonces y=log(x) es normal
- Si x tiene una distribuciónde Poisson, entonces y=sqrt(x) es aprox normal con σ=1

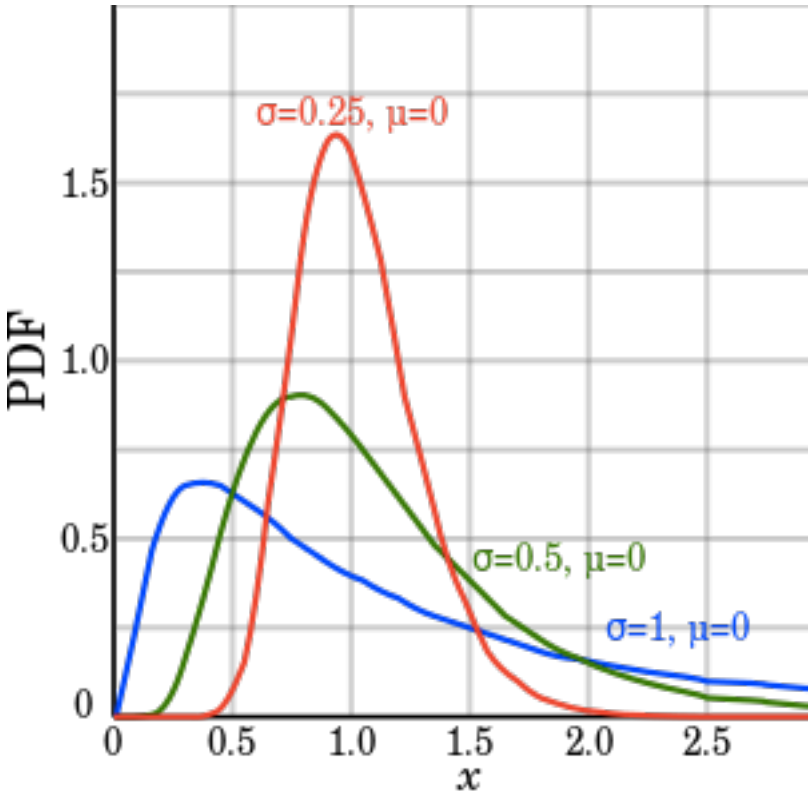
Distribución Log-normal

Tiene solo valores reales positivos

$$\ln \mathcal{N}(x;\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left[-\frac{(\ln x-\mu)^2}{2\sigma^2}\right], \quad x>0$$

$$X = e^{\mu+\sigma Z}$$

Mean	$e^{\mu+\sigma^2/2}$
Median	e^{μ}
Mode	$e^{\mu-\sigma^2}$
Variance	$(e^{\sigma^2}-1)e^{2\mu+\sigma^2}$

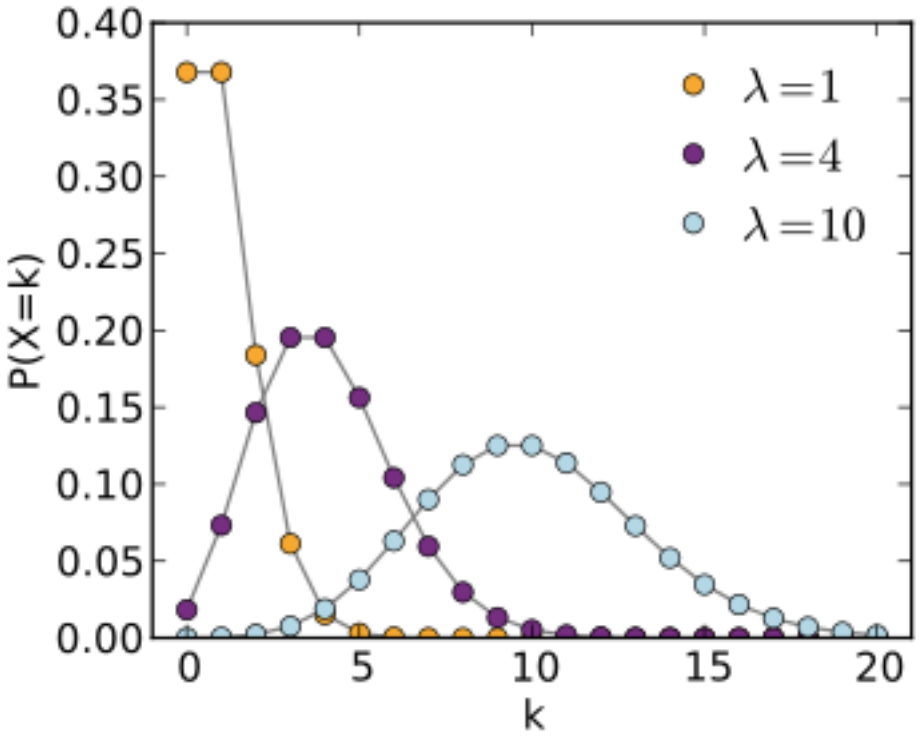


Distribución de Poisson

Probabilidad discreta de que un cierto número de eventos ocurran en un intervalo fijo de tiempo o espacio, si ocurren con una tasa conocida y son independientes

$$F(k,\lambda) = \frac{e^{-\lambda}\lambda^{k}}{k!}$$

k eventos en intervalo
media λ
σ=sqrt(λ)



Distribución Binomial

Probabilidad discreta del número de éxitos, k, en una secuencia de n experimentos independientes con 2 resultados posibles con probabilidad p del éxito

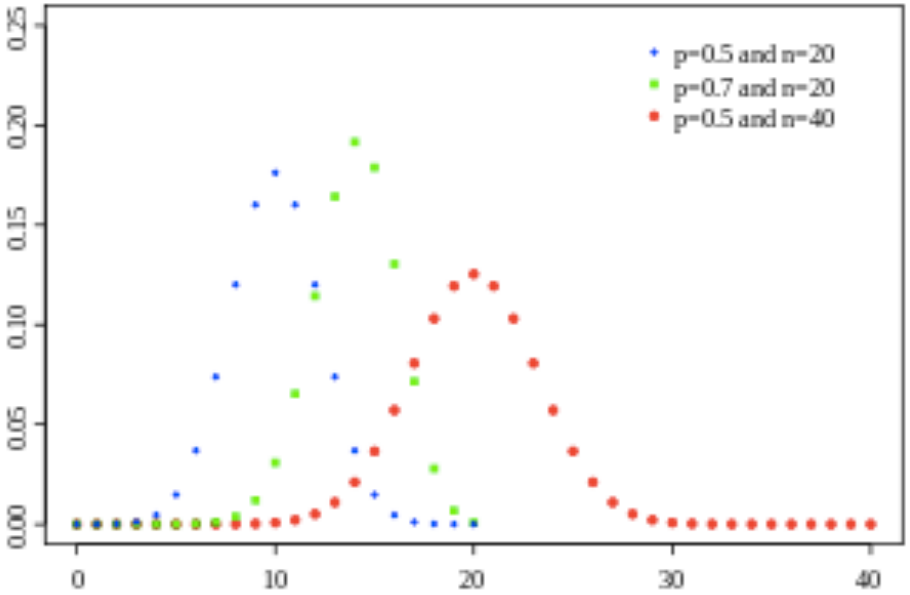
$$f(k;n,p) = \Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$\mathrm{E}[X] = np,$$

$$\mathrm{Var}[X] = np(1-p).$$

Si n es grande se puede aproximar con una distr. normal

$$\mathcal{N}(np, np(1-p))$$



Cuando n es muy grande y p se acerca a 0, entonces la binomial se parece a la poissoniana con np->λ.

Regla: $n \geq 100$ y $np \leq 10$

Distribución Multinomial

Generalización de la binomial.

$$f(x_1, \ldots, x_k; n, p_1, \ldots, p_k) = \Pr(X_1 = x_1 \text{ and } \ldots \text{ and } X_k = x_k)$$

$$= \begin{cases} \dfrac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}, & \text{when } \sum_{i=1}^{k} x_i = n \\ 0 & \text{otherwise,} \end{cases}$$

$$\mathrm{E}(X_i) = np_i.$$

$$\mathrm{var}(X_i) = np_i(1 - p_i).$$

Distribución Exponencial

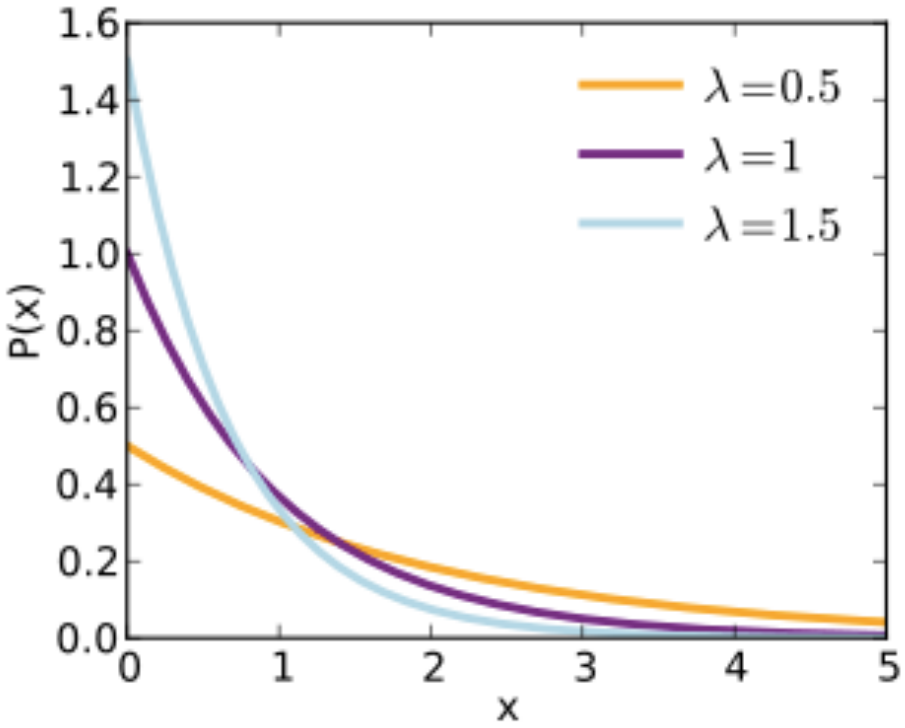
Probabilidad que describe el tiempo entre eventos de un proceso poissoniano

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0, \\ 0 & x < 0. \end{cases}$$

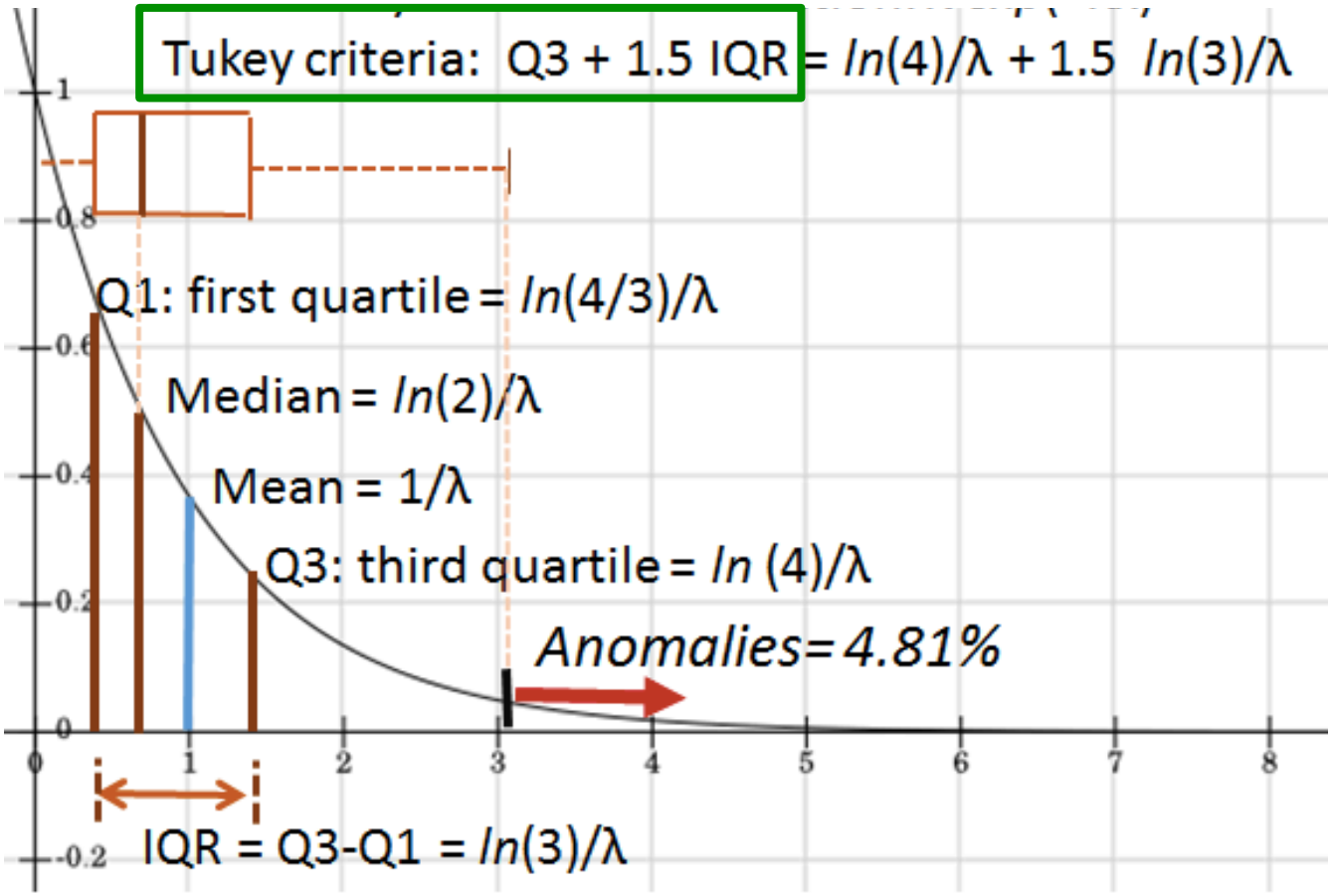
k eventos en intervalo

media λ^{-1} ,

$\sigma=\lambda^{-2}$

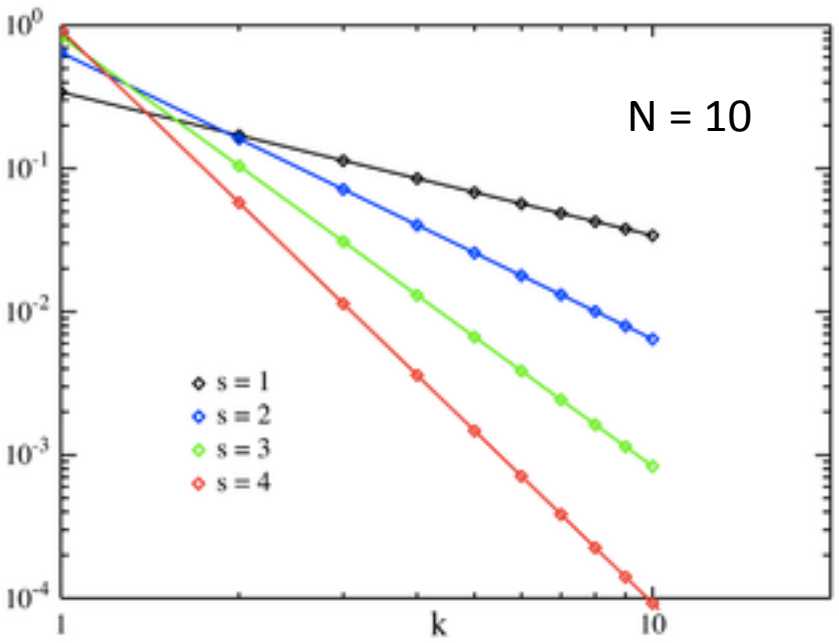


Distribución Exponencial



Otras Distribuciones

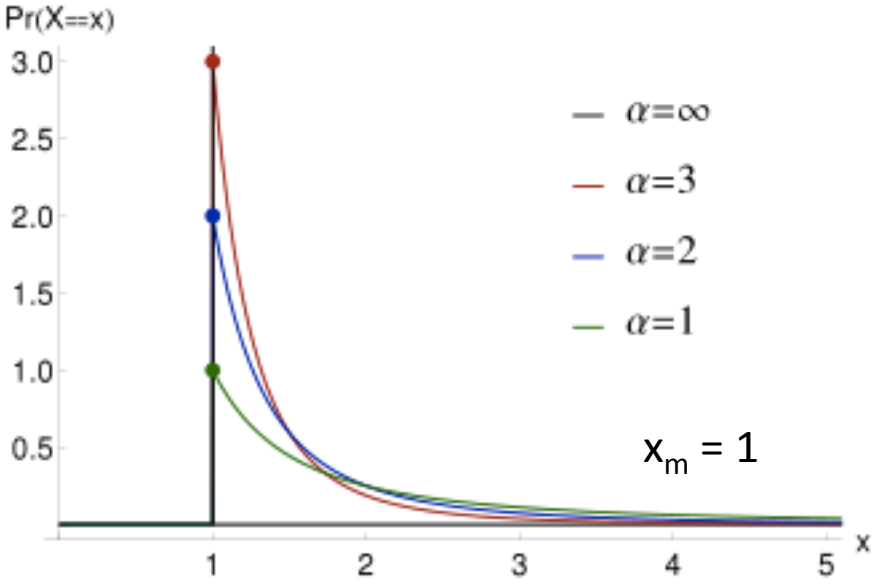
Zipf



$$f(k;s,N) = \frac{1/k^s}{\sum_{n=1}^{N}(1/n^s)}$$

Distribución discreta
N, número de elementos
k, rango s, exponente característico

Pareto



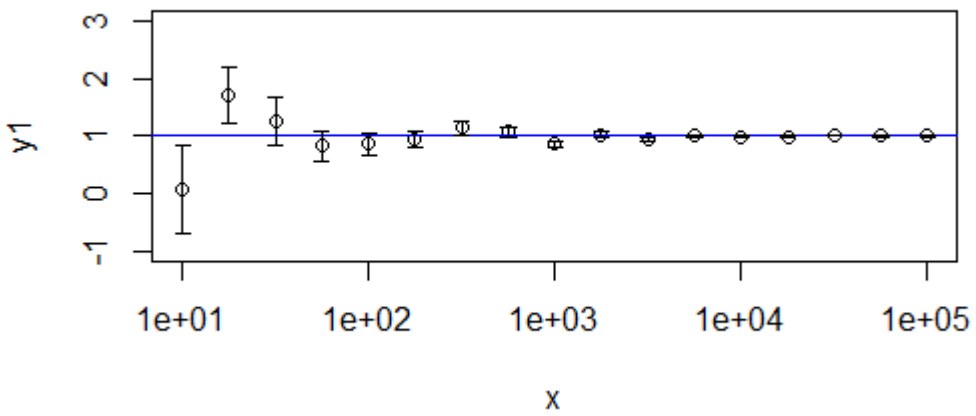
$$f_X(x) = \begin{cases} \frac{\alpha x_m^\alpha}{x^{\alpha+1}} & x \geq x_m, \\ 0 & x < x_m. \end{cases}$$

x_m parámetro de escala
α parámetro de forma

Ejercicio

Para números aleatorios de una función normal (media 1 y σ=2), hacer dos gráficos de dispersión que muestren en el eje x la cantidad de elementos del subconjunto (entre 10 y 10^5, en pasos de 0.25 en el exponente) y en el eje y en un gráfico la media y en el otro la varianza correspondiente. Además añadir una línea horizontal en y=1 e y =2, respectivamente

Media




```
> x<-10^(seq(1,5,0.25))
> z<-sapply( x, function(x) { rnorm(x,1,2) } )
> y1<-sapply(z,mean)
> y2<-sapply(z,var)
> plot(x,y1,log="x", ylim = c(-1,3))
> abline(h=1,col="blue")
```

standard error of the mean (SEM)

$$\sigma_{\bar{X}} = \sqrt{\mathrm{Var}\{\bar{X}\}} = \frac{\sigma}{\sqrt{n}}.$$

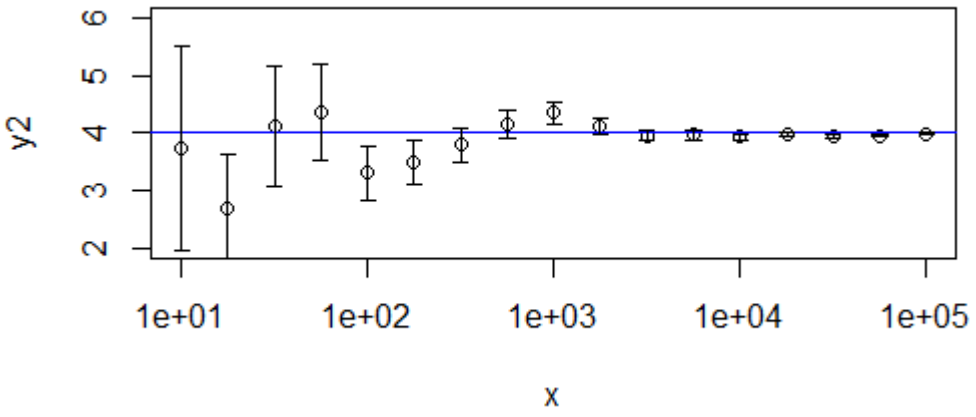
```
>error1<-sqrt(y2/x)
>arrows(x,y1-error1,x,y1+error1,length = 0.05,
angle=90,code=3)
```

Ejercicio

Para números aleatorios de una función normal (media 1 y σ=2), hacer dos gráficos de dispersión que muestren en el eje x la cantidad de elementos del subconjunto (entre 10 y 10^5, en pasos de 0.25 en el exponente) y en el eje y en un gráfico la media y en el otro la varianza correspondiente. Además añadir una línea horizontal en y=1 e y =2, respectivamente

```
> plot(x,y2,log="x", ylim = c(2,6))
> abline(h=4,col="blue")
```

Varianza




```
> error2<-y2*sqrt(2/(x-1))
> arrows(x,y-error2,x,y+error2,length
= 0.05,angle=90,code=3)
```

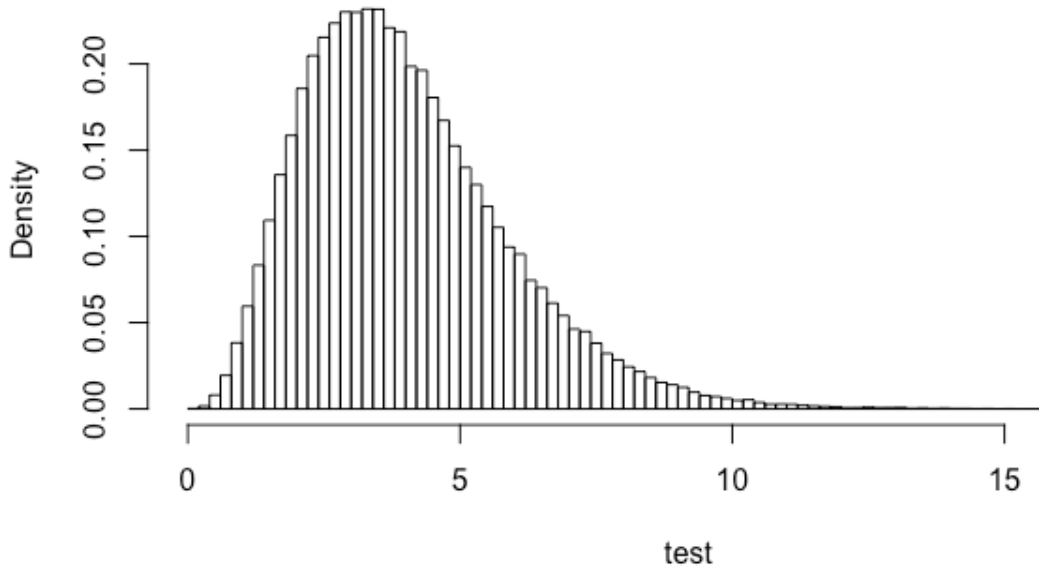
standard error of the variance

$$\sigma_{S^2} = \sqrt{\mathrm{Var}\{S^2\}} = \sigma^2 \sqrt{\frac{2}{n-1}}.$$

Ejercicio

Para el caso anterior hallar mediante simulaciones el intervalo de confianza de la varianza, es decir las barras de error graficadas para las diferentes cantidades de elemento de los grupos hasta 1000 y compararlas con la fórmula analítica.
Primero graficar el histograma de la varianza para el caso de un conjunto de 10, con 100000 simulaciones y hacer un ajuste.

```
vars<-function(n,pres){
 var1<-c()
 for(i in seq(1:pres)) {
  elems<-rnorm(n,1,2)
  var1<-c(var1,var(elems))
 }
 var1    }
```




```
test<-vars(10,100000)
hist(test,breaks = 100, probability = TRUE)
```

Ejercicio

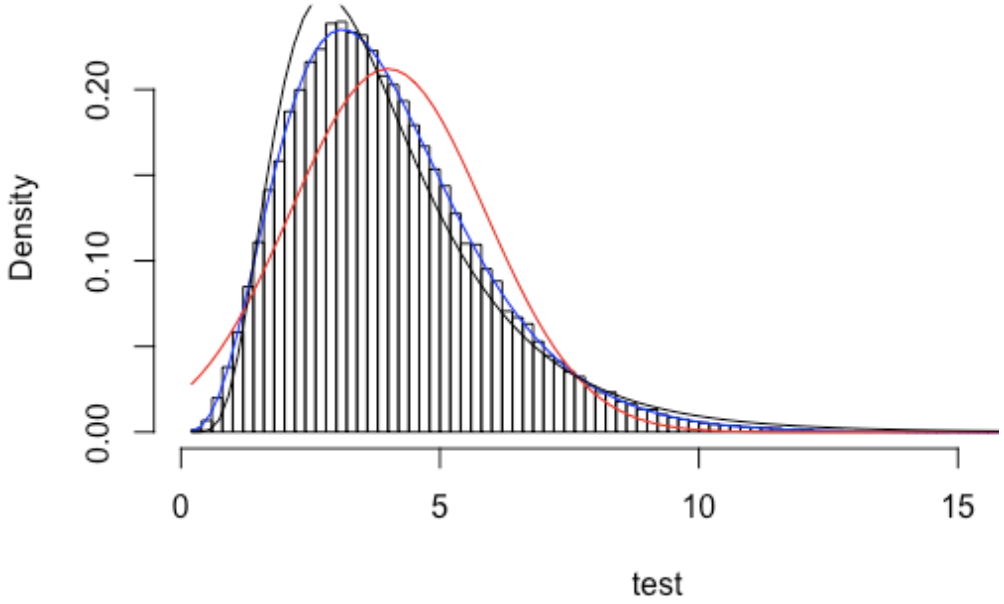
Para el caso anterior hallar mediante simulaciones el intervalo de confianza de la varianza, es decir las barras de error graficadas para las diferentes cantidades de elemento de los grupos hasta 1000 y compararlas con la fórmula analítica.
Primero graficar el histograma de la varianza para el caso de un conjunto de 10, con 100000 simulaciones y hacer un ajuste.

```
library("MASS")

fit1<-fitdistr(test,"gamma")

fit2<-fitdistr(test,"normal")

fit3<-fitdistr(test,"log-normal”)
```




```
curve(dgamma(x,fit1$estimate[1],fit1$estimate[2]),add=TRUE,col="blue")
curve(dnorm(x,fit2$estimate[1],fit2$estimate[2]),add=TRUE,col="red")
curve(dlnorm(x,fit3$estimate[1],fit3$estimate[2]),add=TRUE,col="black")
```

Ejercicio

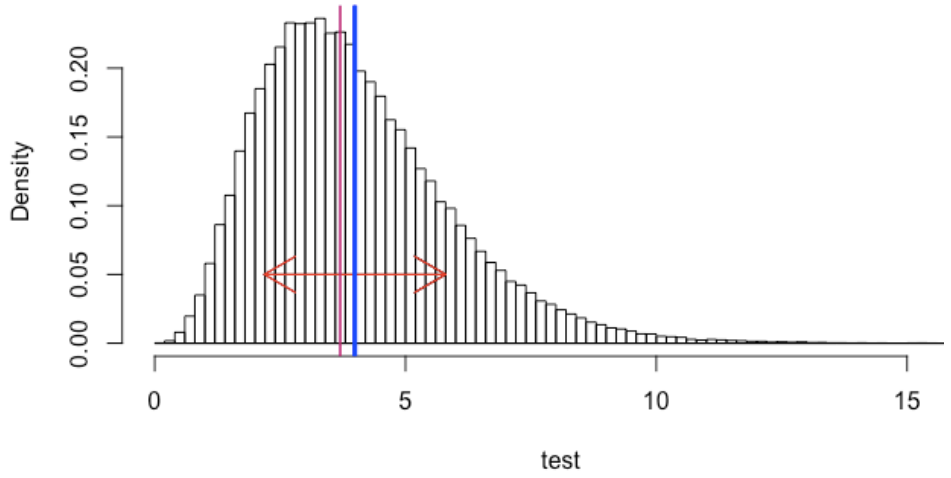
Para el caso anterior hallar mediante simulaciones el intervalo de confianza de la varianza, es decir las barras de error graficadas para las diferentes cantidades de elemento de los grupos hasta 1000 y compararlas con la fórmula analítica.
Primero graficar el histograma de la varianza para el caso de un conjunto de 10, con 100000 simulaciones y hacer un ajuste.

```
quantile(test,probs =
c(0,0.5-0.6827/2,0.5,0.5+0.6827/2,1))
```

```
       0%    15.865%         50%    84.135%        100%
0.1583828  2.1845573  3.6983325  5.8023360 16.1400604
```

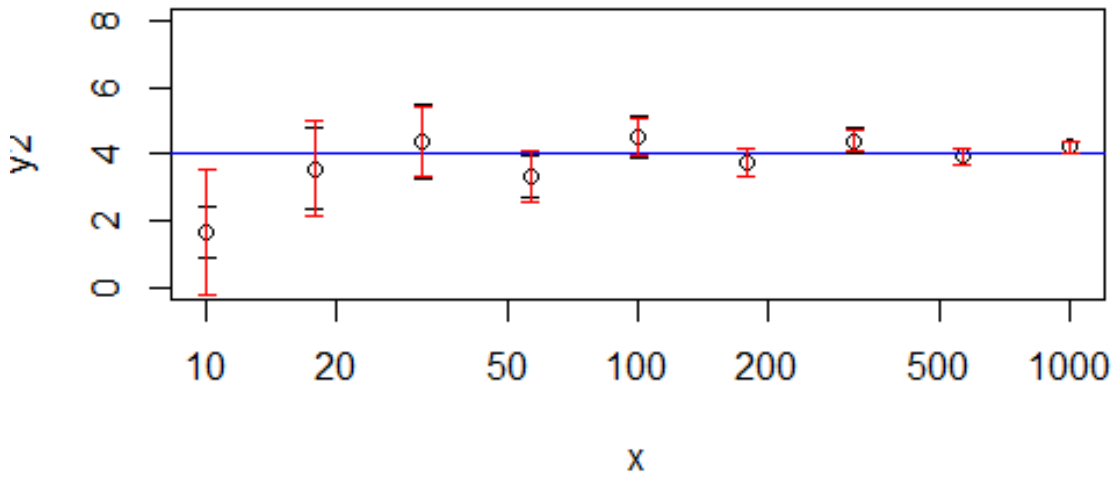
```
arrows(2.1845,0.05,5.8023,0.05,code = 3,col = "red")
abline(v=3.6983,col="red")
abline(v=3.99,col="blue",lw=3)
```

```
> sqrt(var(test))
[1] 1.880598
> (5.802-2.184)/2
[1] 1.809
```



Ejercicio

Para el caso anterior hallar mediante simulaciones el intervalo de confianza de 1σ=68.27% de la varianza, es decir las barras de error graficadas para las diferentes cantidades de elemento de los grupos hasta 1000 y compararlas con la fórmula analítica.



Estimado Analítico
Simulación

```
errorvar<-function(n,pres){
 var1<-c()
 for(i in seq(1:pres)) {
  elems<-rnorm(n,1,2)
  var1<-c(var1,var(elems))
 }
 sqrt(var(var1))    }
```

```
error22<-sapply(x,function(x) {errorvar(x,100000)})
```

```
arrows(x,y2-error22,x,y2+error22,length = 0.05,angle=90,code=3,col = "red")
```

Prueba de hipótesis

Quisiéramos probar la hipótesis H_A, pero es más fácil descartar la **hipótesis nula H_0**
H_0: afirmación que el fenómeno estudiado no produce efectos

"**test statistics**" es una medida de los datos que será grande bajo H_A y pequeña bajo H_0

Se muestra que una hipótesis puede ser válida demostrando la improbabilidad que la opuesta (H_0) sea verdadera

- Un resultado es estadísticamente significativo si puede rechazar H_0.
- Implica que la hipótesis correcta está en el complemento lógico de H_0 , pero no necesariamente será H_A .
- Si H_0 no se puede descartar a un nivel de confianza, no implica que H_0 sea verdadera.

Ejemplo 1

¿Cómo probar que una moneda está trucada?
Se lanza 10 veces y todas son sello.

¿Se puede concluir algo?
Es más fácil decir que la moneda no es justa (H_0) a que está trucada (H_A)

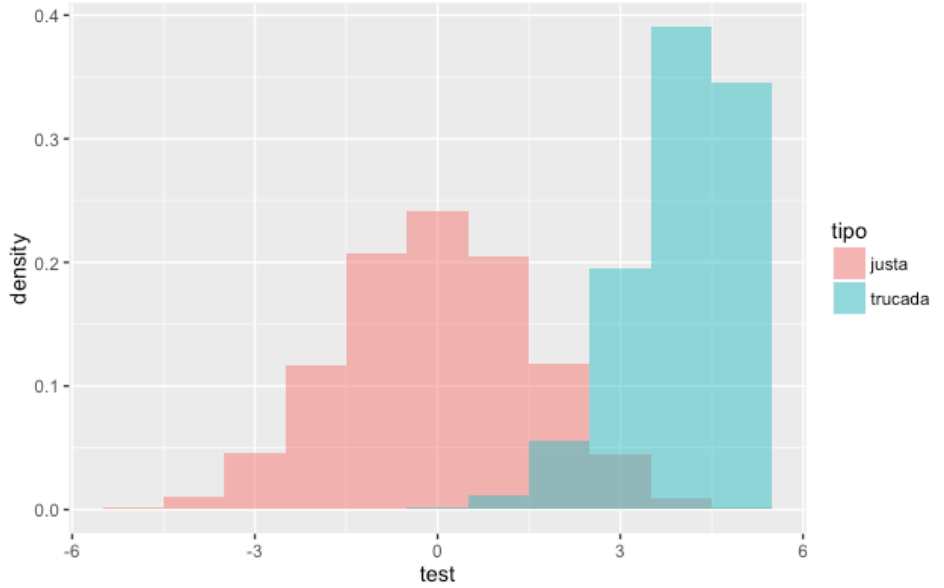
Prueba estadística (s)
(medida de datos que será grande bajo H_A y pequeña bajo H_0):

s = diferencia entre número de sellos y k/2 (prueba con 2 colas)

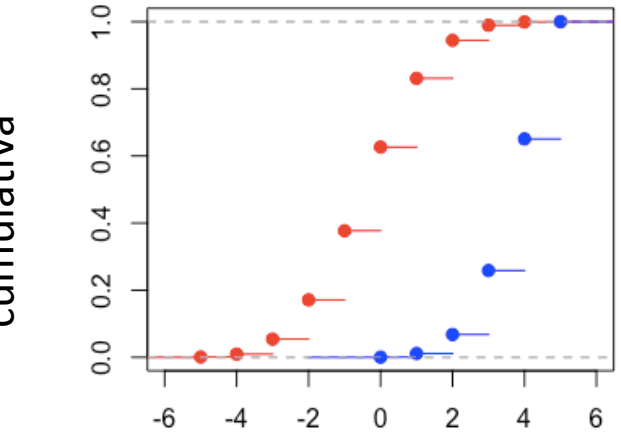
Ejemplo 1

Test = Diferencia entre número de sellos y k/2

```
library(ggplot2)
k<-10
fair<-data.frame(test=rbinom(100000,k,0.5)-k/2)
notfair<-data.frame(test=rbinom(100000,k,0.9)-k/2)
fair$tipo <- 'justa'
notfair$tipo <- 'trucada'
monedas<-rbind(fair,notfair)
ggplot(monedas, aes(test,..density.., fill = tipo)) +
geom_histogram(alpha = 0.5, position = 'identity',
binwidth = 1)
```




```
plot(ecdf(fair$test),col="red")
plot(ecdf(notfair$test),col="blue",add=T)
```




```
h1<-hist(fair$test,breaks=seq(-5,5,1))
h2<-hist(notfair$test,breaks=seq(-5,5,1))
h1$counts[10]/100000      #0.00103
h2$counts[10]/100000      #0.34719
```

Si en un experimento salieron 10 sellos, test=5 tiene una probabilidad = 0.1% si es justa y 34.7% si es trucada.

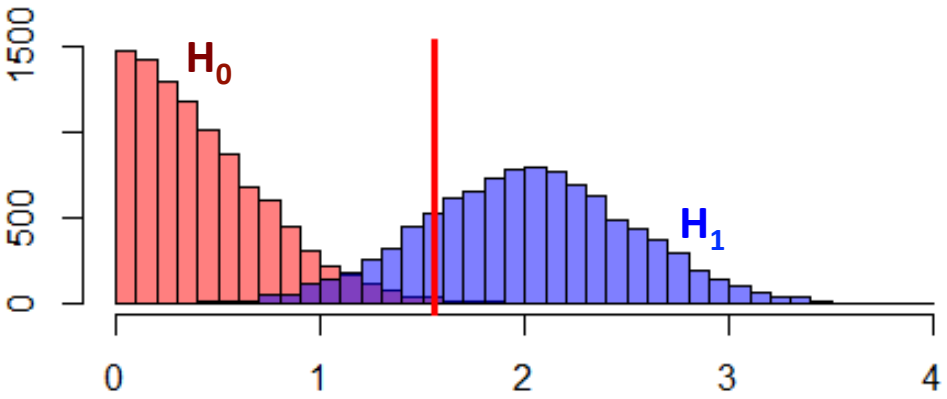
Ejemplo 2

Dos muestras a y b, sacadas de distribuciones normales diversas, son diferentes

Prueba estadística: **s =|media(a)-media(b)|**

donde para H_0 (hipótesis nula) , s sería pequeña pues si las muestras fueran iguales, para la hipótesis que fueran diferentes s sería grande.

```
set.seed(20)
H0<-sapply(rep(60,10000),function(x)
{abs(mean(rnorm(x,13,3))-mean(rnorm(x,13,3)))})
H1<-sapply(rep(60,10000),function(x)
{abs(mean(rnorm(x,12,3.5))-mean(rnorm(x,14,2)))})
hist(H0,breaks = seq(0,4,0.1),col=rgb(1,0,0,0.5),xlim =
c(0,4))
hist(H1,breaks =
seq(0,4,0.1),col=rgb(0,0,1,0.5),add=T)
try<-abs(mean(rnorm(60,12,3.5))-
mean(rnorm(60,14,2)))
abline(v=try,col="red",lw=3)
```



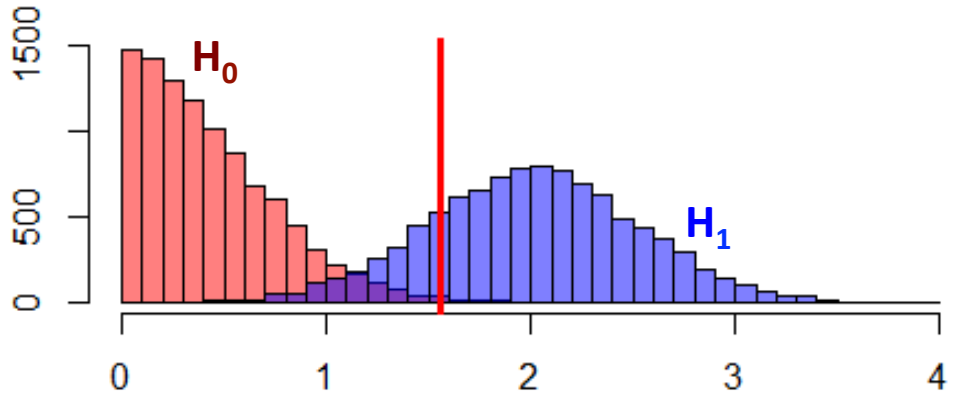
Ejemplo 2

Se **rechaza** H_0 si $\Pr(x > s_{obs} \mid H_0) = p < \alpha$

p: p-value
α: nivel de significancia

la probabilidad que el valor de la prueba estadística sea mayor que el observado debe ser pequeño

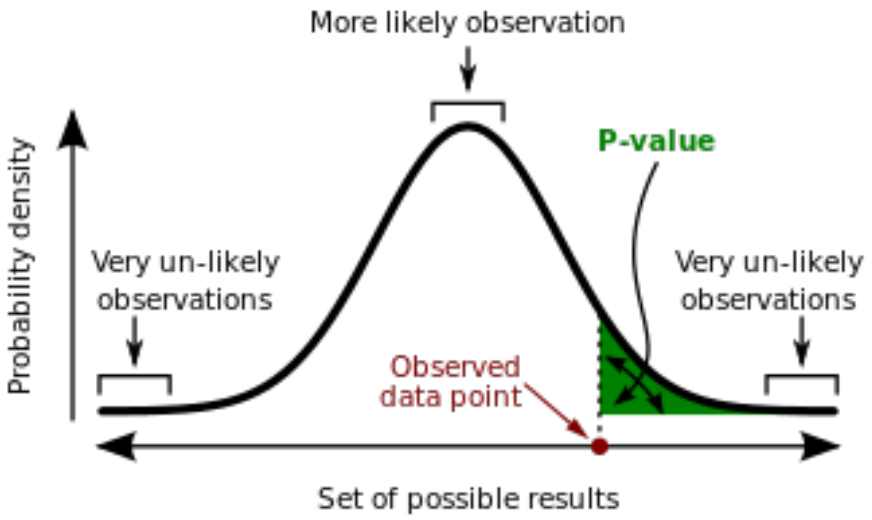
Valor sugerido de α depende del área de investigación:
α=0.05 (2σ)
$\alpha=6\times10^{-7}$ (5σ)



fcum<-ecdf(H0)
p-value<-1-fcum(try)

p-value=0.0051

P-value, significancia



P-value: probabilidad que el resultado observado sea compatible con lo que se quiere probar

$p_{obs} < \alpha$ -> se rechaza la hipótesis nula

Datos observados inconsistentes con H_0

$Pr(obs|H) \neq Pr(H|obs)$

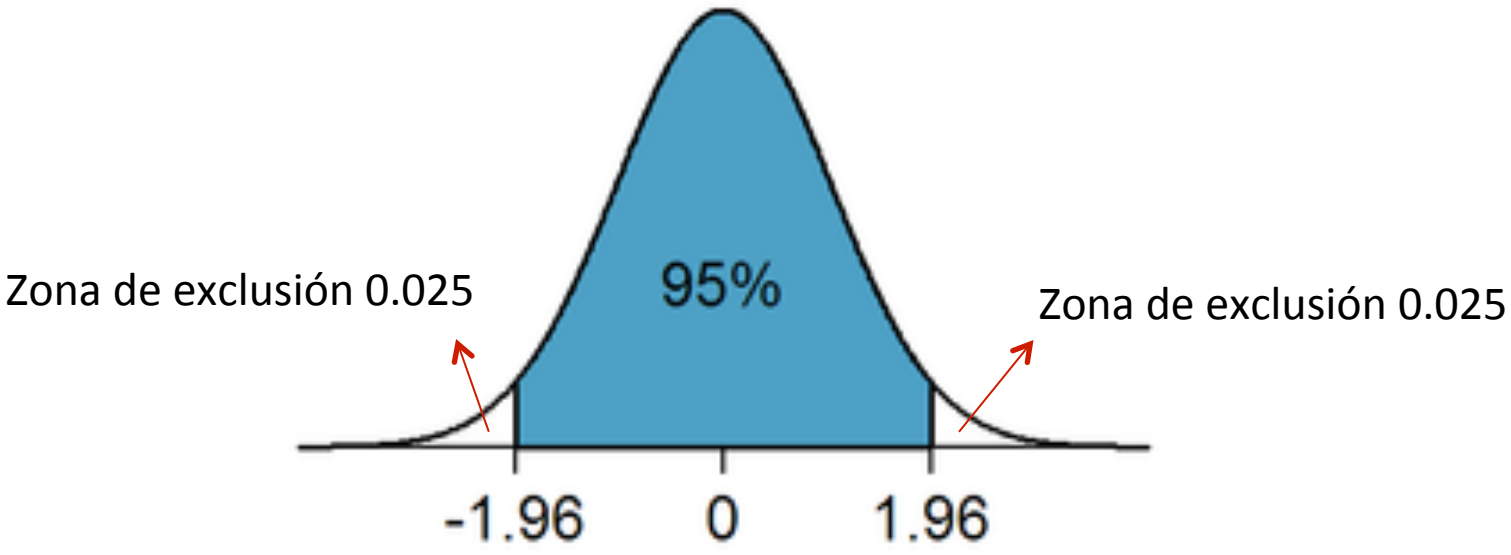
Existen 3 casos: cola derecha, cola izquierda, dos colas

$p = Pr(x \geq s_{obs} \mid H_0)$, cola derecha
$p = Pr(x \leq s_{obs} \mid H_0)$, cola izquierda
$p = \min (Pr(x > s_{obs} \mid H_0), Pr(x \geq s_{obs} \mid H_0))$, dos colas

P-value, significancia

Resultado estadísticamente significativo: cuando p-value<α

dos colas, con α=0.05



Toma de decisiones

	Decisión	
H_0	Mantener H_0	Rechaza H_0
Verdadero	Correcto Pr=1-α **Verdadero positivo**	Error tipo I < α **Falso positivo** Pr(rechazar $H_0 \mid H_0$)=Pr(p≤α$\mid H_0$)=α
Falso	Error tipo II < β **Falso negativo** Pr(mantener $H_0 \mid H_1$)=Pr(p≤β$\mid H_0$)=β	Correcto Pr=1-β **Verdadero negativo**

- Error tipo I: detectar un efecto no presente
- Error tipo II: no lograr detectar un efecto presente

Única forma de reducir ambos errores es incrementar tamaño de muestra.

Prueba de Hipótesis

En un test de clasificación binario se tiene:

- **Sensibilidad**: mide proporción de verdaderos positivos identificados correctamente

= verdaderos positivos/ (verdaderos positivos + falsos negativos)

TP/(TP+FN)

- **Especificidad**: mide proporción de verdaderos negativos identificados correctamente

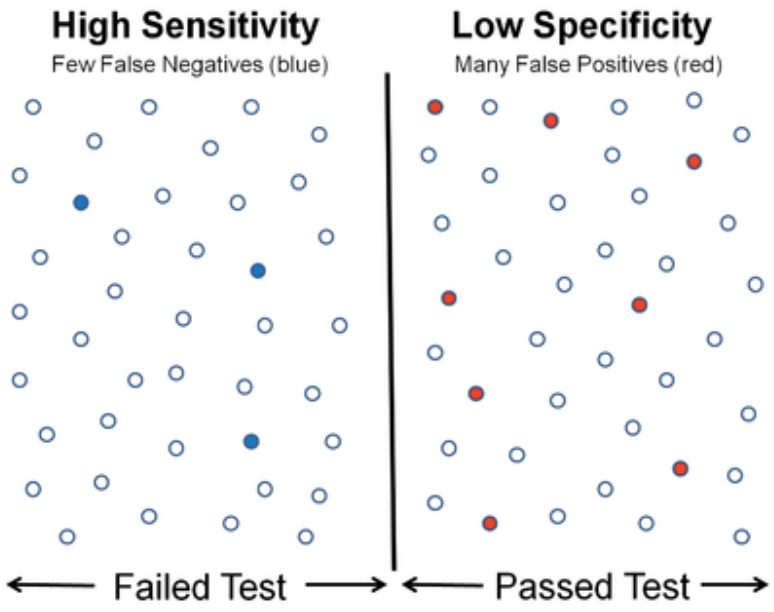
= verdaderos negativos/ (verdaderos negativos + falsos positivos)

TN/(TN+FP)

Predicción perfecta: 100% sensible y 100% específica -> imposible siempre hay un error.

poder estadístico= 1-β

= Pr(rechazar $H_0 | H_1$ es verdadera)



Ejemplo

Tabla de contingencia o matriz de confusión

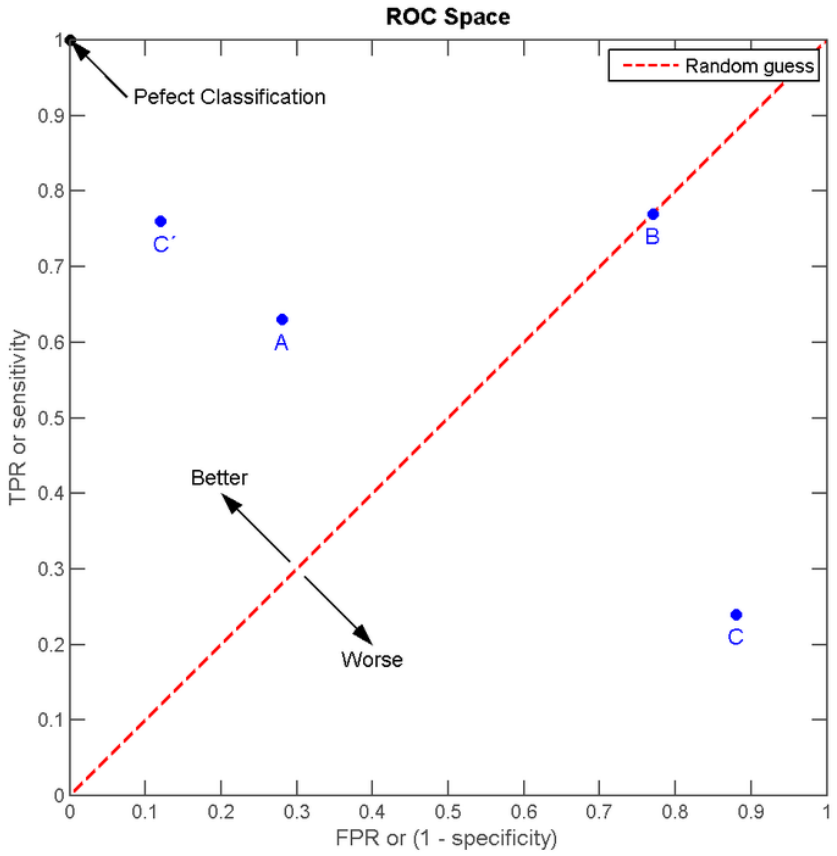
	Condition positive	Condition negative	
Test outcome positive	**True positive** (TP) = 20	**False positive** (FP) = 180	Positive predictive value = TP / (TP + FP) = 20 / (20 + 180) = **10%**
Test outcome negative	**False negative** (FN) = 10	**True negative** (TN) = 1820	Negative predictive value = TN / (FN + TN) = 1820 / (10 + 1820) ≈ **99.5%**
	Sensitivity = TP / (TP + FN) = 20 / (20 + 10) ≈ **67%**	**Specificity** = TN / (FP + TN) = 1820 / (180 + 1820) = **91%**	

Muestra Total 2030

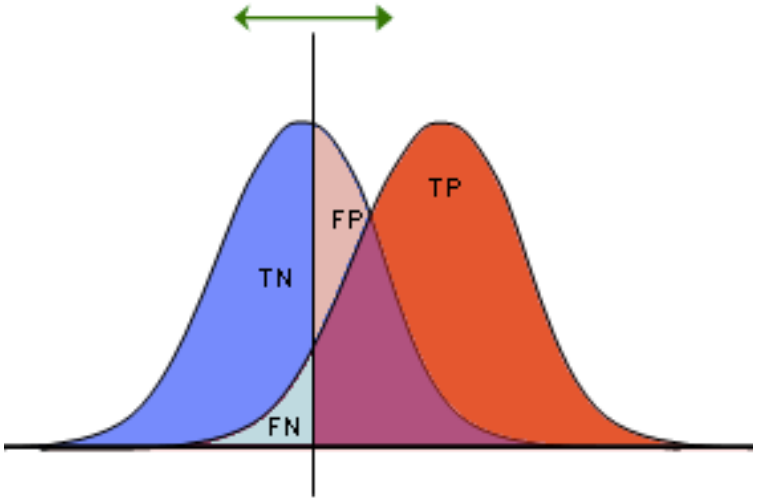
ROC

Curva ROC: **Receiver Operating Characteristic**

Sensibilidad versus (1 – especificidad) para un sistema clasificador binario variando el umbral de discriminación.



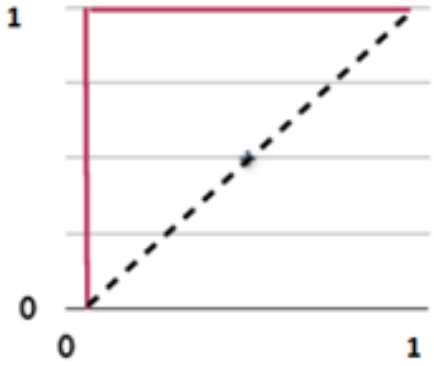
Herramienta para seleccionar modelos óptimos y descartar subóptimos



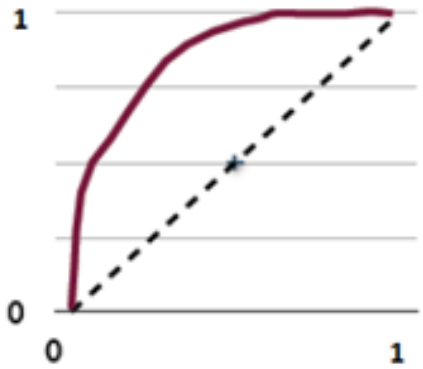
ROC

Curva ROC: **Receiver Operating Characteristic**

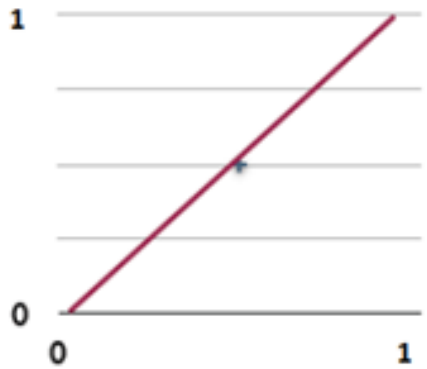
Prueba perfecta



Prueba buena



Prueba mala



Área bajo la curva [0,5; 1] (**AUC**: area under curve) permite comparar bondad de la prueba:

- 1 diagnóstico perfecto
- 0,5 sin capacidad discriminatoria

Ejercicio

Realizar una curva ROC para el ejemplo 2

H_0	True	False
Reject	FP=12	TN=99
Keep	TP=89	FN=1

corte, s=0.8

sen=0.98 ,esp=0.89

H_0	True	False
Reject	FP=3	TN=97
Keep	TP=98	FN=3

corte, s=1.1

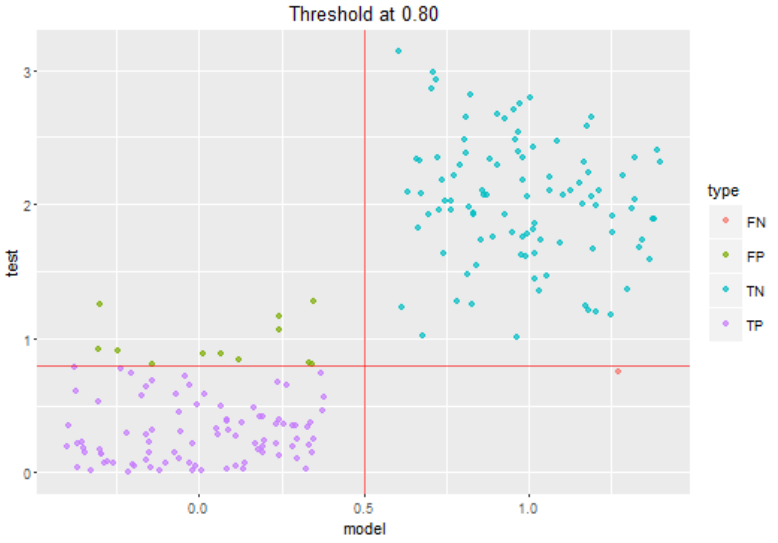
sen=0.97 ,esp=0.97

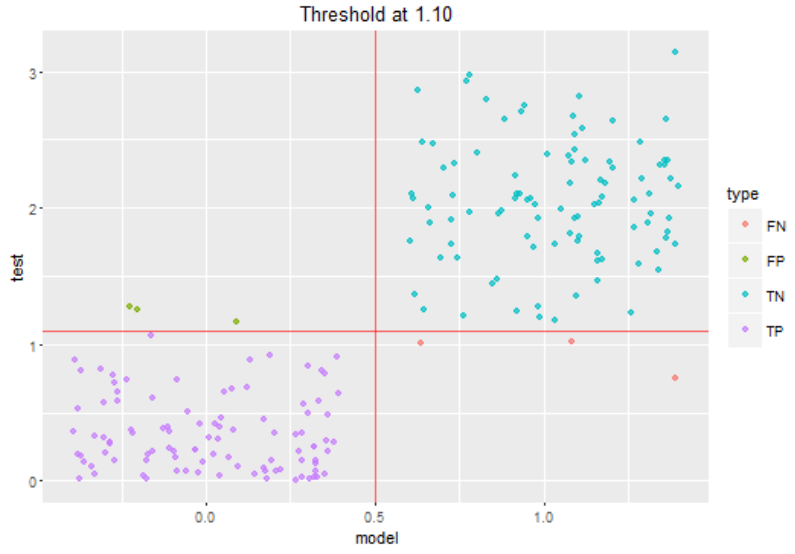
```
plot_test_H_distribution <- function(df, threshold) {
  v <- rep(NA, nrow(df))
 v <- ifelse(df$test < threshold & df$model == 0, "TP", v)
  v <- ifelse(df$test >= threshold & df$model == 0, "FP", v)
  v <- ifelse(df$test < threshold & df$model == 1, "FN", v)
  v <- ifelse(df$test >= threshold & df$model == 1, "TN", v)

df$pred_type <- v
print(table(df$pred_type))

  ggplot(data=df, aes(x=model, y=test)) +
    geom_jitter(aes(color=pred_type), alpha=0.6) +
    geom_hline(yintercept=threshold, color="red", alpha=0.6) +
    geom_vline(xintercept=0.5, color="red", alpha=0.6) +
    scale_color_discrete(name = "type") +
    labs(title=sprintf("Threshold at %.2f", threshold))
}
```

fuente



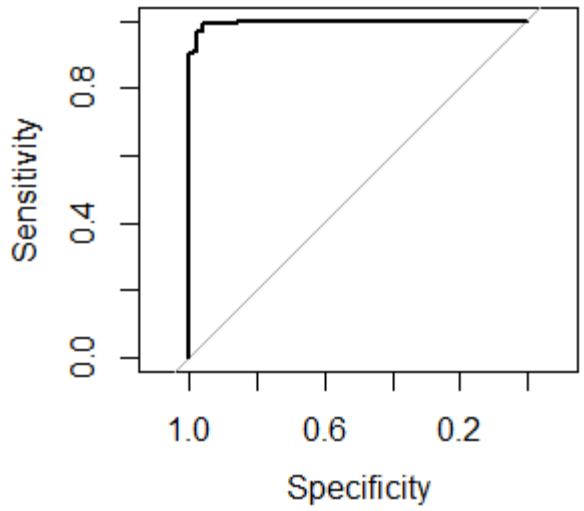


Ejercicio

pROC

Realizar una curva ROC para el ejemplo 2

```
install.packages("pROC")
library(pROC)
df<-data.frame("model"=0,"test"=0)
for(i in seq(100)) {df<-rbind(df, c(1,abs(mean(rnorm(60,12,3.5))-mean(rnorm(60,14,2)))))}
for(i in seq(100)) {df<-rbind(df, c(0,abs(mean(rnorm(60,13,3))-mean(rnorm(60,13,3)))))}
roc(df$model,df$test,plot = T)
```

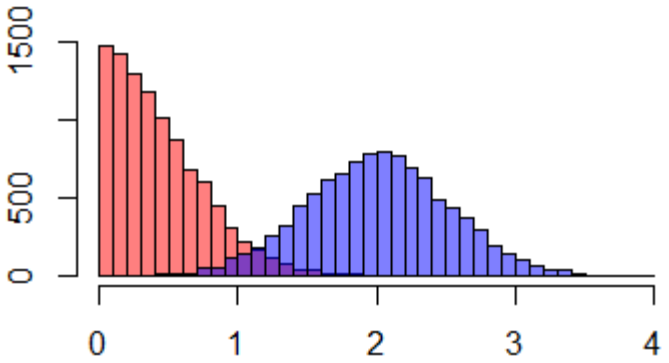


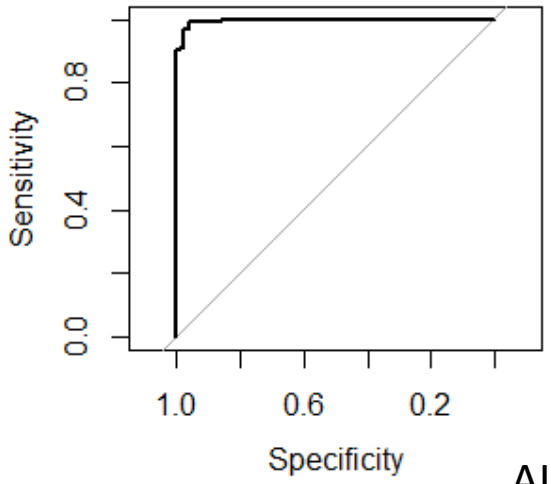

```
Call:
roc.default(response = df$model, predictor = df$test, plot = T)

Data: df$test in 101 controls (df$model 0) < 100 cases (df$model 1).
Area under the curve: 0.9965
```

Ejercicio

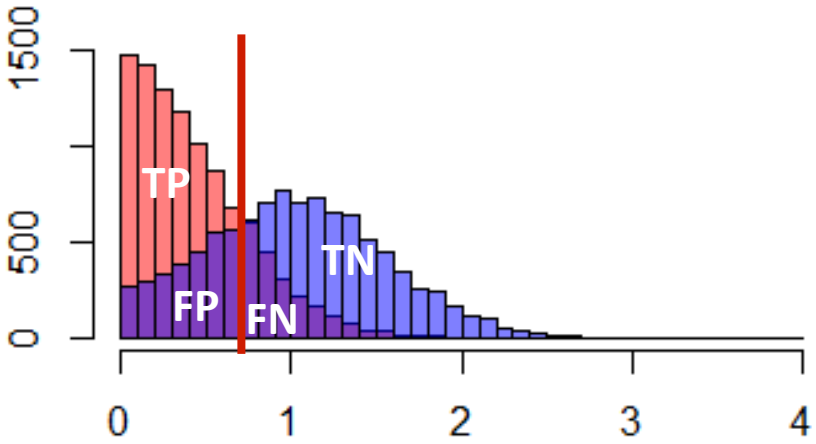
Si H1 viene de $|\mu_{rnorm(60,12,3.5)}-\mu_{rnorm(60,14,2)}|$

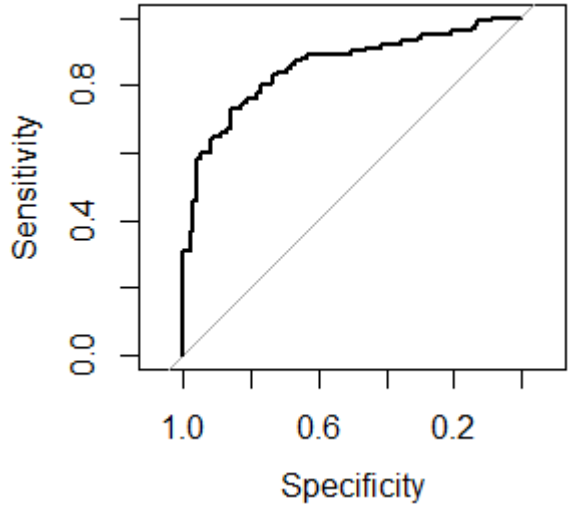




AUC: 0.9965

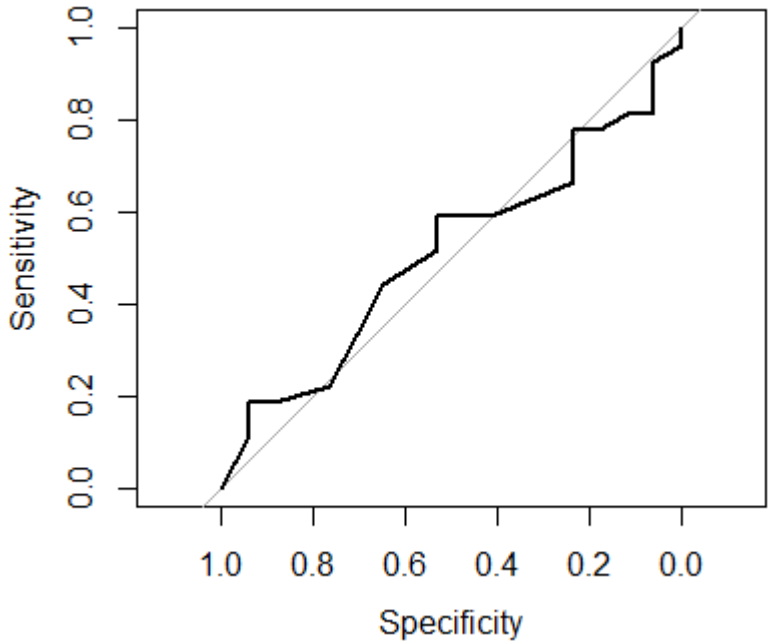
Si H1 viene de $|\mu_{rnorm(60,12,3)}-\mu_{rnorm(60,13,3)}|$





AUC: 0.8547

Ejemplo TEA



AUC: 0.5109

```
df<-read.csv("TEA_Promedios.csv")
df$bien <- df$Promedio>11
library(pROC)
roc(df$bien,df$TEA,plot = T)
```

TEA	Promedio	bien
2.3	6.8	FALSE
3.0	8.4	FALSE
2.3	8.5	FALSE
1.6	8.8	FALSE
2.5	8.9	FALSE
2.0	9.1	FALSE

44 observaciones

Distribución T-Student

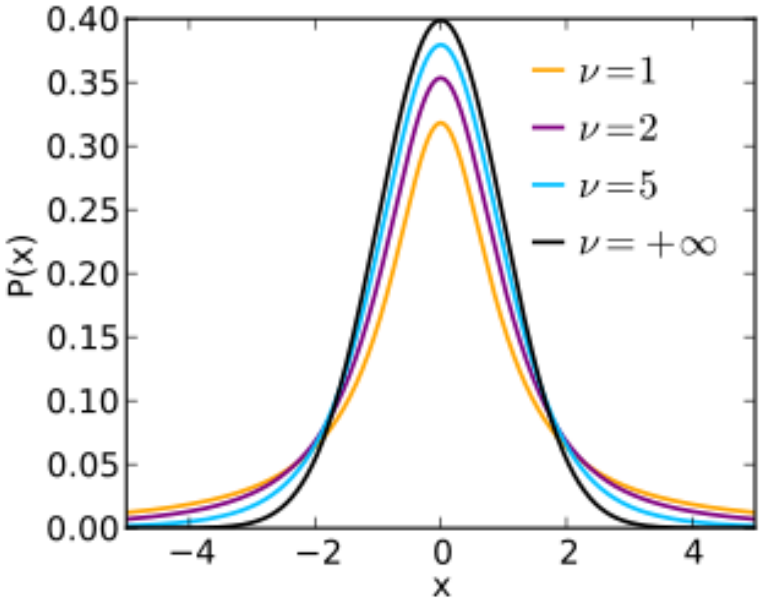
Distribución t-Student o T

Al estimar la media de una población normal cuando la muestra es pequeña y no se conoce σ

Student era el pseudónimo de su creador William Sealy Gosset

$$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

ν=ndf
grados de libertad



Para una muestra de n observaciones de una distribución normal, se tienen una distribución t con ndf=n-1 para la diferencia entre la media poblacional y la de la muestra, dividida entre la desviación estándar de la muestra, todo multiplicado por el término de normalización sqrt(n)

Prueba: T-test

Prueba t de Student comparación de dos grupos

n<30

Cuando la población tiene distribución normal pero la muestra es pequeña y por tanto su estadístico no está normalmente distribuido (se utiliza una estimación de σ)
Entonces el estadístico tiene una distribución t de Student, si la hipótesis nula es cierta.

T-test statistic (conjunto único): $$T \equiv \frac{Z}{\sqrt{V/\nu}} = \left(\overline{X}_n - \mu\right)\frac{\sqrt{n}}{S_n}$$

Z distr. Normal con media 0 y varianza 1
V tiene distr. χ^2 con v=n-1 grados de libertad

media de las medias muestrales

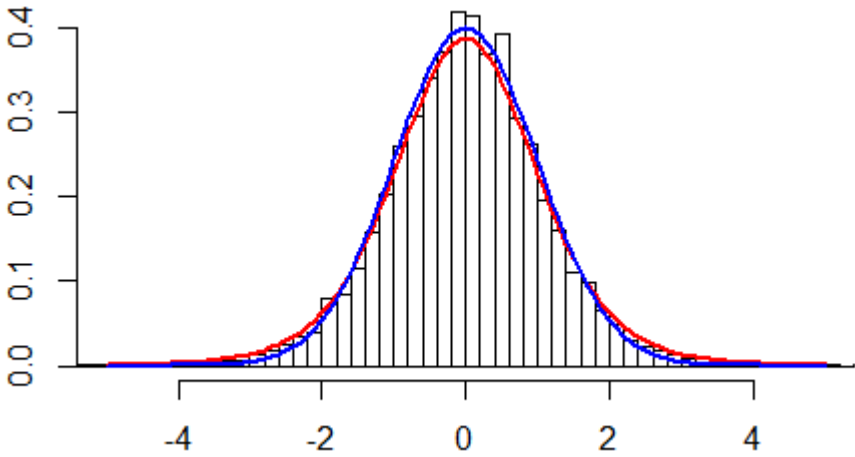
Usos:

- Estimar la significancia estadística de la diferencia entre las medias de dos muestras.
- Construcción de intervalos de confianza para la diferencia entre medias de 2 poblaciones
- Análisis de regresión lineal

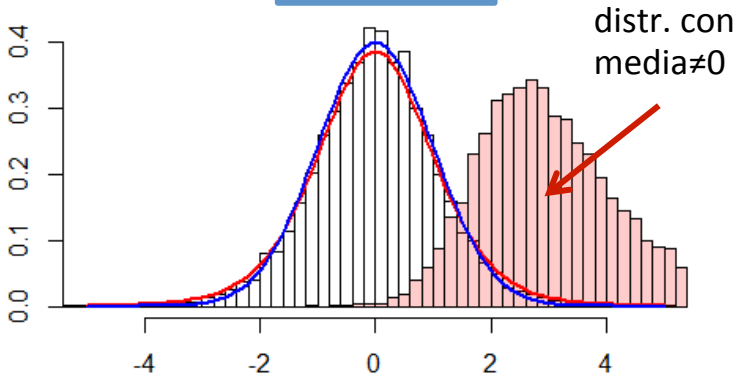
Prueba: T-test de una muestra

Probemos si la media de una distribución es cero (H_0)

```
set.seed(20)
n=9
ndf=n-1
x<-c(n+1,rep(n,10000))
z<-sapply( x, function(x) { rnorm(x,0,1) } )
y1<-sapply(z,mean)
y2<-sapply(z,var)
y3<-y1/sqrt(y2/ndf)
hist(y3,breaks = seq(-10,10,0.2),probability = T,xlim = c(-5,5))
library("MASS")
fit1<-fitdistr(y3,"t")
curve(dt(x,fit1$estimate[3]),add=TRUE,col="red",lw=2)
curve(dnorm(x,0,1),add=TRUE,col="blue",lw=2)
```



$$t = \frac{\bar{x}}{\sigma}\sqrt{n}$$



Prueba: T-test de dos muestras

Dadas dos muestras x_1 y x_2 de tamaño n_1 y n_2

T-test statistic :

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sigma_{\bar{x}_1 - \bar{x}_2}} \qquad \sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}$$

between-subjects test

Para determinar si las medias de dos muestras son iguales

Alternative Hypothesis	Rejection Region
$H_a: \mu_1 \neq \mu_2$	$\lvert T \rvert > t_{1-\alpha/2,\nu}$
$H_a: \mu_1 > \mu_2$	$T > t_{1-\alpha,\nu}$
$H_a: \mu_1 < \mu_2$	$T < t_{\alpha,\nu}$

Distribución χ^2

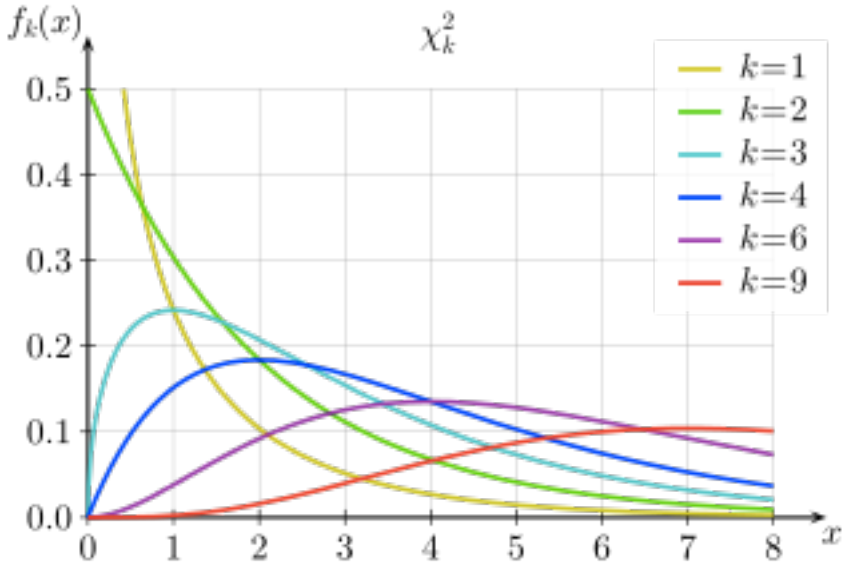
Distribución χ^2 con k grados de libertad: distr. de suma de cuadrados de k variables aleatoria con distribución normal estándar.

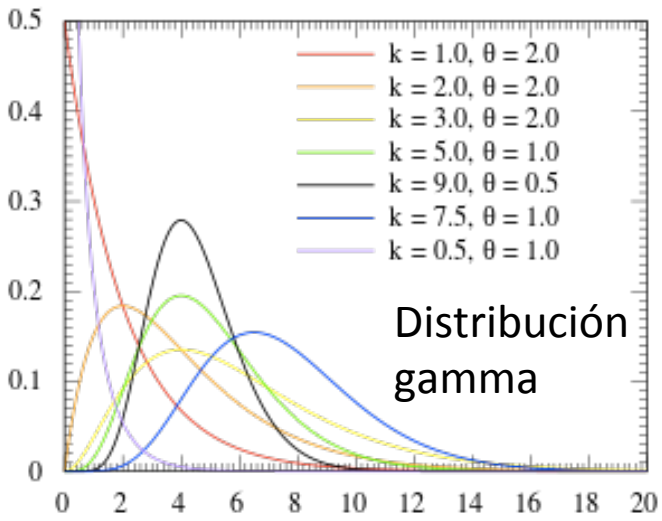
$$Q = \sum_{i=1}^{k} Z_i^2,$$

$$f(x;\, k) = \begin{cases} \frac{x^{(k/2-1)} e^{-x/2}}{2^{k/2}\Gamma\left(\frac{k}{2}\right)}, & x > 0; \\ 0, & \text{otherwise.} \end{cases}$$

$$\Gamma(n) = (n-1)!$$

Caso especial de distribución gamma.





Distribución gamma

Prueba χ^2

Pearson's χ^2 test

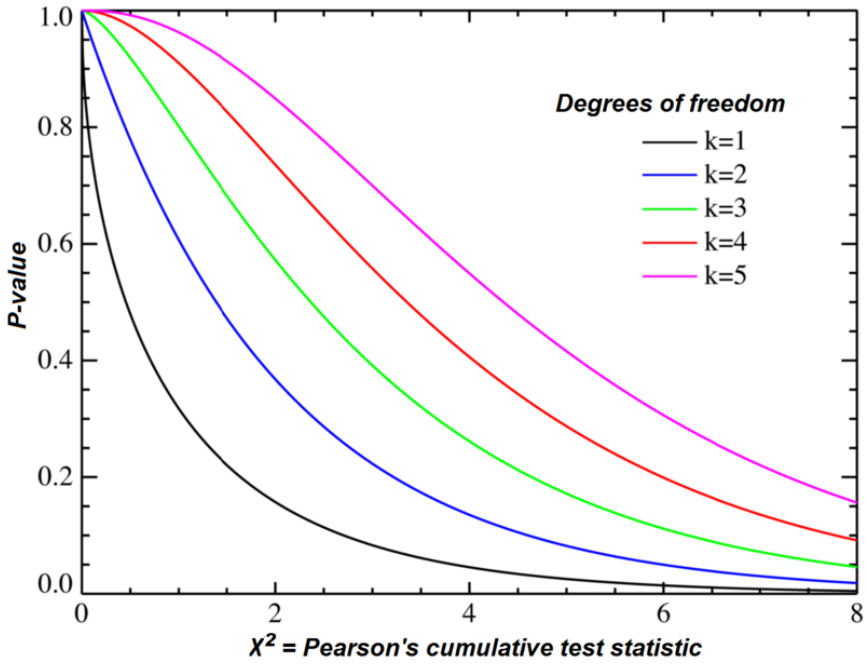
χ^2 prueba si la observación (conteo) es consistente con los datos (modelo)
Se usa para rechazar H_0 (datos independientes)

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} = N \sum_{i=1}^{n} \frac{(O_i/N - p_i)^2}{p_i}$$

$$E_i = N p_i$$

O_i cuentas observadas del tipo i
E_i valor esperado (teórico, modelo)
N número total de observaciones
p_i fracción del tipo i en la población

Si H_0 es verdadera χ^2 sigue una distr. χ^2



Dos tipos de comparación:
-**bondad del ajuste**: si lo observado difiere de la distribución teórica
-**prueba de independencia** dos variables observadas

Prueba χ^2

χ^2 se usa para calcular p-value comparando χ^2 con la distribución χ^2 con v ndf

ndf : número de grados de libertad = n – r,
n: número de categorías (tipos, celdas)
r: reducción en grados de libertad, r=par+1 con par= parámetros de distribución ajustada

Ejemplos:

- Distribución uniforme: para N observaciones divididas en n categorías:
  E_i=N/n, r=1, -> ndf=n-r=n-1 (cuentas observadas están restringidas a sumar N, se pierde un grado de libertad)
- Distribución normal: con 2 parámetros, μ y σ, -> r=3
- Distribución de Poisson: 1 parámetro, λ, -> r=2

Name	Statistic
chi-squared distribution	$\sum_{i=1}^{k} \left(\frac{X_i - \mu_i}{\sigma_i}\right)^2$
noncentral chi-squared distribution	$\sum_{i=1}^{k} \left(\frac{X_i}{\sigma_i}\right)^2$

X^2, ndf, p-value

Degrees of freedom (df)	χ^2 value[18]										
1	0.004	0.02	0.06	0.15	0.46	1.07	1.64	2.71	3.84	6.64	10.83
2	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.60	5.99	9.21	13.82
3	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.82	11.34	16.27
4	0.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	9.49	13.28	18.47
5	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	15.09	20.52
6	1.63	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	16.81	22.46
7	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	18.48	24.32
8	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	20.09	26.12
9	3.32	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	21.67	27.88
10	3.94	4.87	6.18	7.27	9.34	11.78	13.44	15.99	18.31	23.21	29.59
P value (Probability)	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001

Bondad de ajuste

$$\chi^2 = \sum \frac{(O-E)^2}{\sigma^2}$$

Definición usada cuando se tiene un estimado del error (σ) de la medida, asumiendo que los errores tienen un distribución normal

$$\chi^2_{\text{red}} = \frac{\chi^2}{\nu} = \frac{1}{\nu}\sum \frac{(O-E)^2}{\sigma^2}$$

X^2 reducido

ν número de grados de libertad = N-n
(N observaciones y n parámetros ajustados)

Regla (válida cuando la varianza es conocida a priori y no estimada de los datos) :
$\chi^2 >> 1$ mal ajuste del modelo o error subestimado
$\chi^2 = 1$ indica que el ajuste entre observaciones y modelo está de acuerdo dentro del error
$\chi^2 < 1$ sobre ajuste del modelo: modelo ajustando inapropiadamente el ruido o el error ha sido sobrestimado

Prueba de independencia

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{i,j}-E_{i,j})^2}{E_{i,j}}$$

$$= N\sum_{i,j} p_{i\cdot}p_{\cdot j}\left(\frac{(O_{i,j}/N)-p_{i\cdot}p_{\cdot j}}{p_{i\cdot}p_{\cdot j}}\right)^2$$

$$E_{i,j} = Np_{i\cdot}p_{\cdot j},$$

$$p_{\cdot j} = \frac{O_{\cdot j}}{N} = \frac{\sum_{i=1}^{r} O_{i,j}}{N}$$

$$p_{i\cdot} = \frac{O_{i\cdot}}{N} = \sum_{j=1}^{c}\frac{O_{i,j}}{N},$$

Observación consiste en valores de 2 resultados.
H_0 = ocurrencia de dichos eventos es estadísticamente independiente

Se coloca cada observación en una tabla de contingencia, arreglo bi-dimensional (r=row, c=column), de acuerdo a los resultados de las 2 variables.

Handedness / Gender	Right handed	Left handed	Total
Male	43	9	52
Female	44	4	48
Total	87	13	100

Prueba de independencia

```
example <- matrix(c(43,9,44,4),ncol=2,byrow=TRUE)
colnames(example) <- c("RH","LH")
rownames(example) <- c("Male","Female")
example <- as.table(example)
library(MASS)
chisq.test(example)
```

```
        Pearson's Chi-squared test with Yates' continuity
        correction

data:  example
X-squared = 1.0725, df = 1, p-value = 0.3004
```

Handed-ness / Gender	Right handed	Left handed	Total
Male	43	9	52
Female	44	4	48
Total	87	13	100

Prueba exacta de Fisher

Para solo cuentas en dos diferentes condiciones

	Count1(X)	Count2(X)
X=0	a	b
X=1	c	d

El test exacto de Fisher con n=a+b+c+d , será:

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+b)!\ (c+d)!\ (a+c)!\ (b+d)!}{a!\ \ b!\ \ c!\ \ d!\ \ n!}$$

Da directamente la probabilidad, no es una estadística

Pruebas paramétricas

Pruebas paramétricas asumen que los datos están normalmente distribuidos y que las muestras son independientes y tienen la misma distribución.

Siempre verificar que los datos satisfacen estas suposiciones.

Tomar en cuenta:

- Outliers : gran efecto si se usa estimados de varianzas
- Valores correlacionados como muestras (repetir medida en el mismo sujeto)
- Distribuciones asimétricas: dan resultados inválidos

Pruebas no paramétricas

No hacen suposiciones sobre distribución de datos y pueden ser usados en juegos de datos arbitrarios

Prueba de Kolmogorov-Smirnov (K-S)

Verificar si dos distribuciones (continuas o discretas) son similares

Estadístico: distancia máxima entre funciones cumulativas de las distribuciones

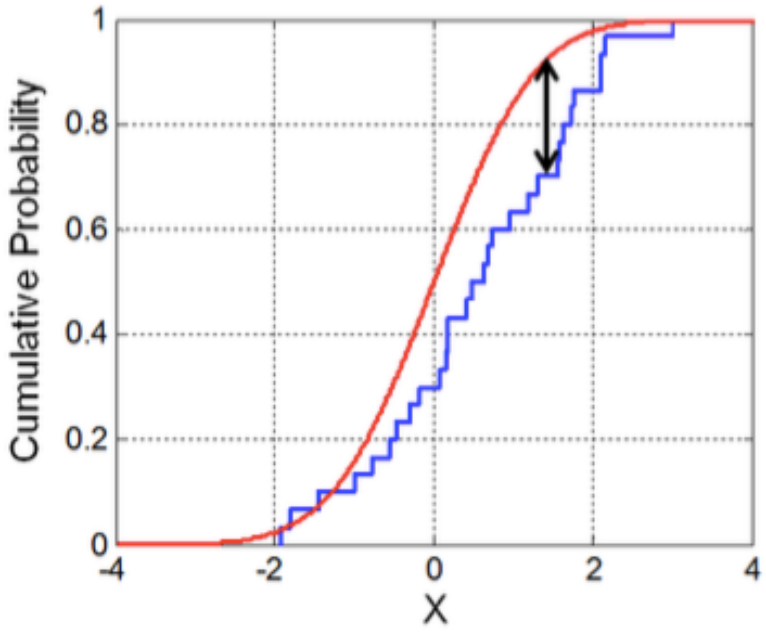
Prueba de un lado: distribución observada (histograma) comparada contra distribución de referencia

$$D_n = \sup_x |F_n(x) - F(x)|$$

F=CDF

Prueba de dos lados: dos distribuciones de observaciones son comparadas.

$$D_{n,n'} = \sup_x |F_{1,n}(x) - F_{2,n'}(x)|,$$

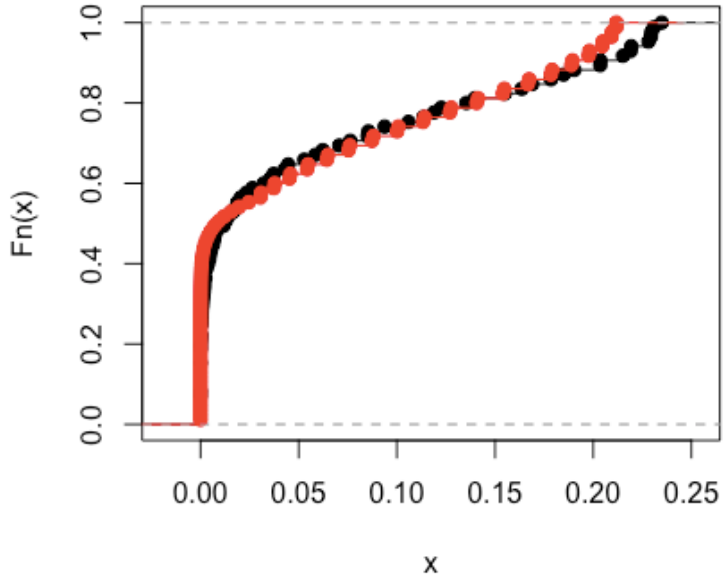


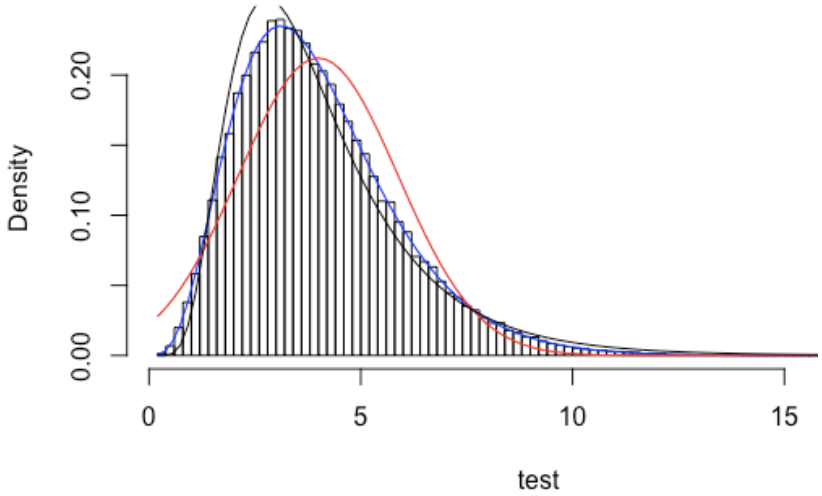
Kolmogorov-Smirnov

Del histograma del ejemplo en slide 23

```
observations<-histo$density
model1<-sapply(histo$mids,function(x){dgamma(x,fit1$estimate[1],fit1$estimate[2])})
model2<-sapply(histo$mids,function(x){dnorm(x,fit2$estimate[1],fit2$estimate[2])})
model3<-sapply(histo$mids,function(x){dlnorm(x,fit3$estimate[1],fit3$estimate[2])})
```

```
plot(ecdf(observations),col=”black")
> plot(ecdf(model2), add = TRUE,col=“red")
```





Kolmogorov-Smirnov

Del histograma del ejemplo en slide 23

```
observations<-histo$density
model1<-sapply(histo$mids,function(x){dgamma(x,fit1$estimate[1],fit1$estimate[2])})
model2<-sapply(histo$mids,function(x){dnorm(x,fit2$estimate[1],fit2$estimate[2])})
model3<-sapply(histo$mids,function(x){dlnorm(x,fit3$estimate[1],fit3$estimate[2])})
```

ks.test(observations, model1)

```
        Two-sample Kolmogorov-Smirnov test

data:  observations and model1
D = 0.070588, p-value = 0.9839
alternative hypothesis: two-sided
```

ks.test(observations, model2)

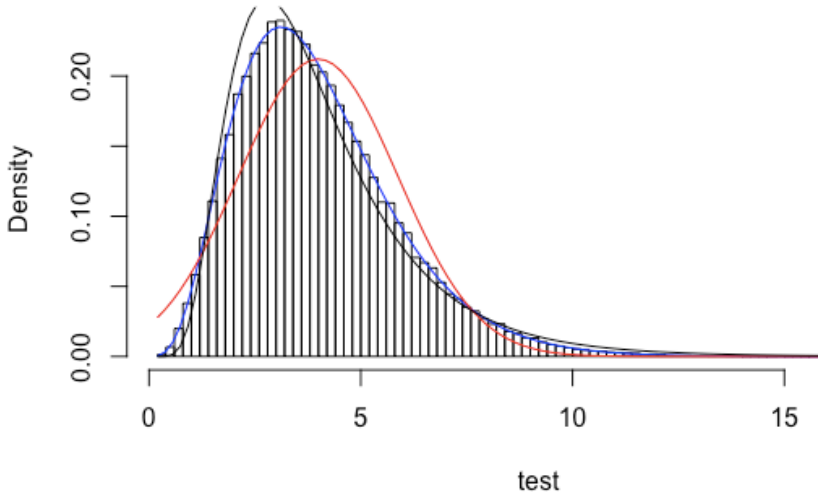
```
        Two-sample Kolmogorov-Smirnov test

data:  observations and model2
D = 0.25882, p-value = 0.006731
alternative hypothesis: two-sided
```

ks.test(observations, model3)

```
        Two-sample Kolmogorov-Smirnov test

data:  observations and model3
D = 0.22353, p-value = 0.02861
alternative hypothesis: two-sided
```



Pruebas de normalidad

http://www.r-bloggers.com/normality-tests-don%E2%80%99t-do-what-you-think-they-do/