Transformación de Datos

Transformación Raíz Cuadrada Si las observaciones tiene una distribución de Poisson debe usarse $\sqrt{y_{ij}}$ o $\sqrt{1+y_{ij}}$

Transformación Logarítmica (para respuestas positivas) Si los datos tiene una distribución Lognormal ($\ln\left(Y_{ij}\right) \sim$ Normal), entonces la transformación es logarítmica $\ln\left(Y_{ij}\right)$.

Transformación Seno Inverso Para datos binomiales expresado en fracciones se debe usar la transformación seno inverso $\text{sen}^{-1}\sqrt{y_{ij}}$

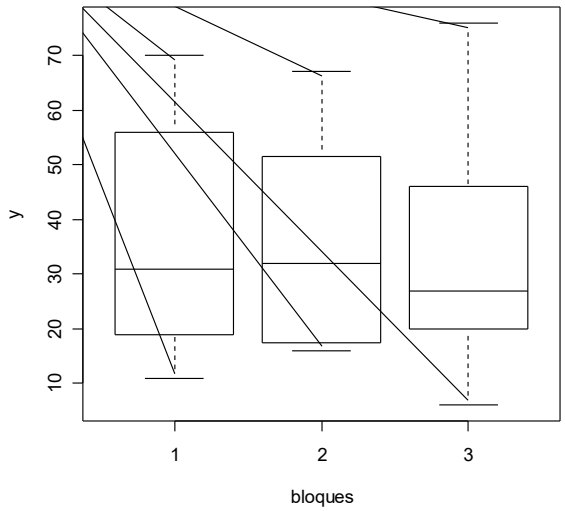
Ejemplos:
1.- Auhtry(1948) presenta los siguientes datos sobre la simbiosis del cruce de Medicago sativa(53) – M. Falcata(50) cruzados con la cepa B. Los datos son porcentajes de plantas con nódulos de un total de 20 por celda. El experimento fue realizado como un diseño de bloques completos al azar.

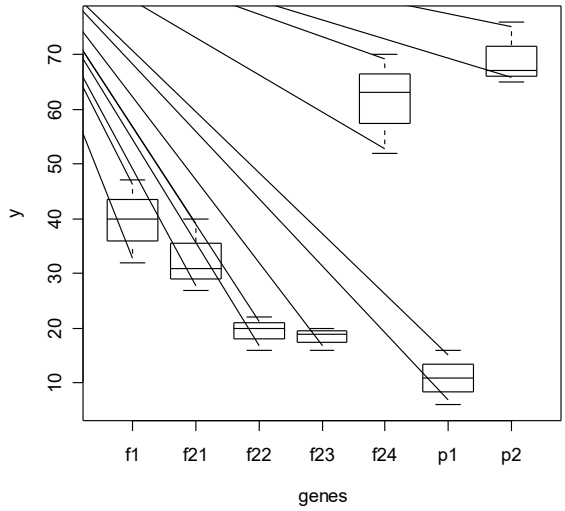
Bloques	Padres		F_1	Lotes de F_2 de cada F_1			
	53	50	53×50	114-1	114-2	114-3	114-4
1	11	65	47	31	22	16	70
2	16	67	32	40	16	19	63
3	6	76	40	27	20	20	52

Como los datos están expresados en porcentajes se origina de una distribución Binomial, Por tanto la transformación más adecuada es arco seno inverso. Para realizar estas transformaciones se divide primero entre 100 y luego se aplica la transformación $\text{sen}^{-1}\sqrt{y_{ij}}$. El cual es realizado con el paquete R que transforma a radianes.

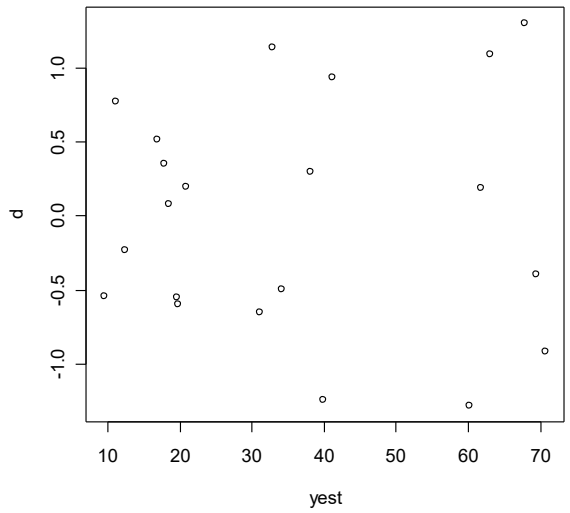
```
> simb<-read.table("simbiosis.txt"header=T)
> y<-simb[,1]
> bloques<-as.factor(simb[,3])
> genes<-simb[,2]
> modb<-lm(y~bloques+genes)
> anva<-aov(modb)
> summary(anva)
> summary(anva)
            Df  Sum Sq  Mean Sq  F value  Pr(>F)
bloques      2    31.7    15.9    0.3962  0.6814
genes        6  9028.0  1504.7   37.5943  4.257e-07 ***
Residuals   12   480.3    40.0
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
```

```
> plot(y~bloques+genes)
```






```
> modb<-lm(y~bloques+genes)
> e<-residuals(modb)
> cme<-deviance(modb)/12
> d<-e/sqrt(cme)
> yest<-fitted(modb)
> plot(yest,d)
```




```
> y1<-asin((y/100)^.5)
> modg1<-lm(y1~bloques+genes)
> anva1<-aov(modg1)
```

```
> summary(anva1)
            Df  Sum Sq   Mean Sq  F value  Pr(>F)
bloques      2  0.00439  0.00219   0.4287  0.661
genes        6  1.07779  0.17963  35.1061  6.246e-07 ***
Residuals   12  0.06140  0.00512
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
```

Transformaciones para estabilizar Variancia

Sea $E[Y] = \mu$ la media de Y: Supóngase que la desviación estándar es proporcional a alguna potencia de la media de Y, tal que

$$\sigma_Y \propto \mu^{\alpha}$$

Se desea determinar la transformación de Y que produzca una variancia constante. Se supone que la transformación es una potencia de los datos originales, Esto es

$$Y^* = Y^{\lambda}$$

Entonces se puede demostrar que:

$$\sigma_Y \propto Y^{\lambda+\alpha-1}$$

Se puede observar claramente que para que los datos transformado es una constante si $\lambda = 1-\alpha$. En la siguiente tabla se resumen algunas de las transformaciones más usadas para estabilizar la variancia. Nótese en este caso si $\lambda = 0$, la transformación es logarítmica:

Relación entre σ_Y y μ	α	$\lambda = 1-\alpha$	Transformación
$\sigma_Y \propto$ constante	0	1	Ninguna
$\sigma_Y \propto \mu^{1/2}$	1/2	1/2	Raíz cuadrada
$\sigma_Y \propto \mu$	1	0	Logarítmica
$\sigma_Y \propto \mu^{3/2}$	3/2	-1/2	Recíproca de la Raíz cuadrada
$\sigma_Y \propto \mu^2$	2	-1	Recíproca

En muchas situaciones de diseño experimental en las que se usan réplicas, α puede estimarse empíricamente a partir de los datos. Puesto que la combinación del i-ésimo de los tratamientos $\sigma_{y_i} \propto \mu_i^{\alpha} = \theta \mu_i^{\alpha}$, donde θ es una constante de proporcionalidad, puede tomarse logaritmo natural para obtener:

$$\ln \sigma_{y_i} = \ln \theta + \alpha \ln \mu_i$$

Por lo tanto, una gráfica de $\ln \sigma_{y_i}$ contra $\ln \mu_i$ sería una línea recta con pendiente α. Puesto como no se conoce σ_{y_i} y μ_i puede sustituirse estimaciones razonables como la desviación estándar (S_i) y la media ($\bar{y}_i$) de las observaciones para el tratamiento i en lugar de σ_{y_i} y μ_i, respectivamente

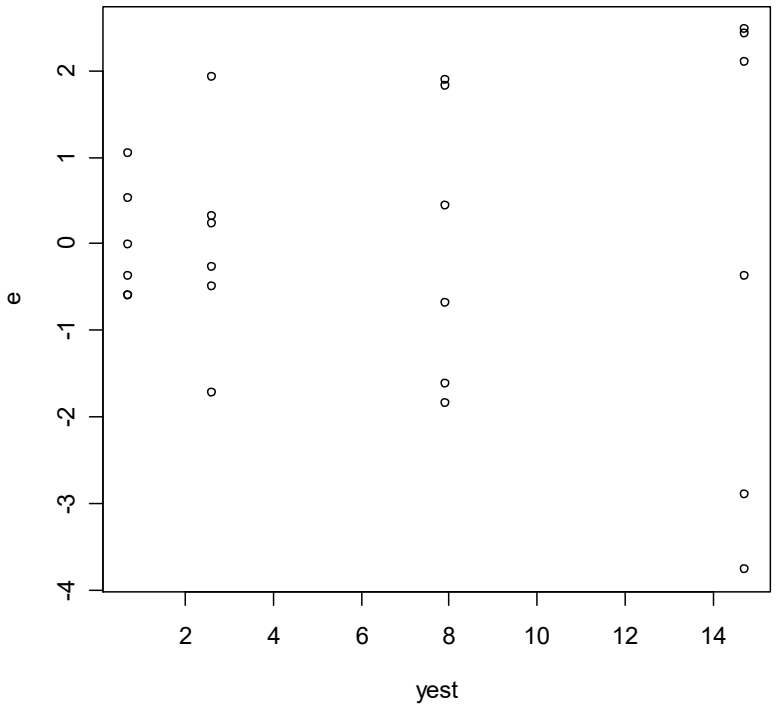
Ejemplo: Un ingeniero civil está interesado en determinar si cuatro métodos diferentes para estimar la frecuencia de las inundaciones producen estimaciones equivalentes de la descarga pico cuando se aplican a la misma cuenca. Cada procedimiento se usa seis veces en la cuenca, y los datos de las descargas resultantes (en pies cúbicos por segundo) se muestran en la siguiente tabla:

Método de Estimación	Observaciones					
1	0.34	0.12	1.23	0.70	1.75	0.12
2	0.91	2.94	2.14	2.36	2.86	4.55
3	6.31	8.37	9.75	6.09	9.82	7.24
4	17.15	11.82	10.95	17.20	14.35	16.82

```
> descarga<-read.table("descarga.txt",header=T) # crear cada uno la data txt
> y<-descarga[,1]
> metodo<-as.factor(descarga[,2])
> mod1<-lm(y~metodo)
> anova(mod1)
```

```
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
metodo     3 708.35  236.12  76.067 4.111e-11 ***
Residuals 20  62.08    3.10
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

> e<-residuals(mod1)
> yest<-predict(mod1)
> plot(yest,e)



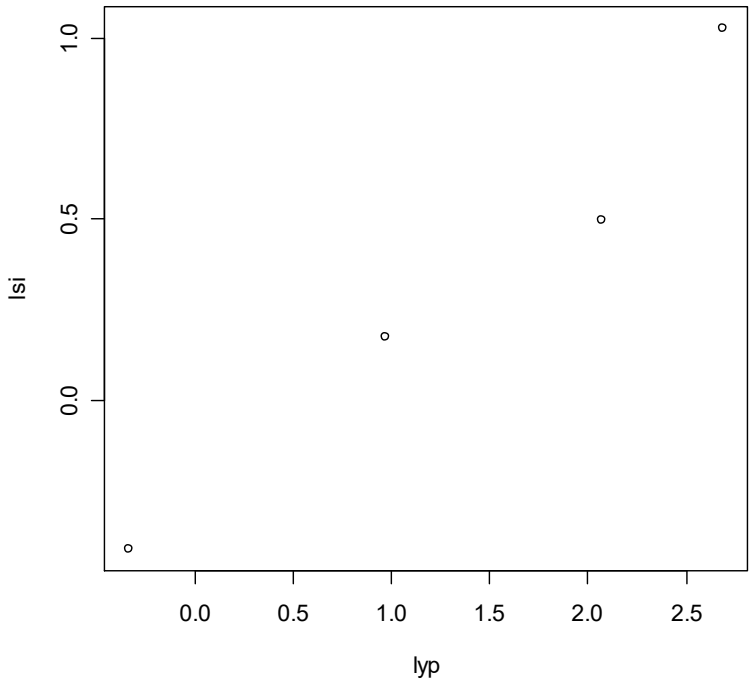
> ae<-abs(e)

```
> anova(lm(ae~metodo))
Analysis of Variance Table

Response: ae
          Df  Sum Sq Mean Sq F value   Pr(>F)
metodo     3 11.5397  3.8466   6.406 0.003222 **
Residuals 20 12.0093  0.6005
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> bartlett.test(y~metodo)
```

```
        Bartlett test of homogeneity of variances

data:  y by metodo
Bartlett's K-squared = 8.9958, df = 3, p-value = 0.02935
```

Entonces no existe homogeneidad de variancias en cuanto a las descargas entre los cuatro métodos de evaluación.

```
> yp<-tapply(y,metodo,mean)
> si<-tapply(y,metodo,sd)
> lyp<-log(yp)
> lsi<-log(si)
> plot(lyp,lsi)
```

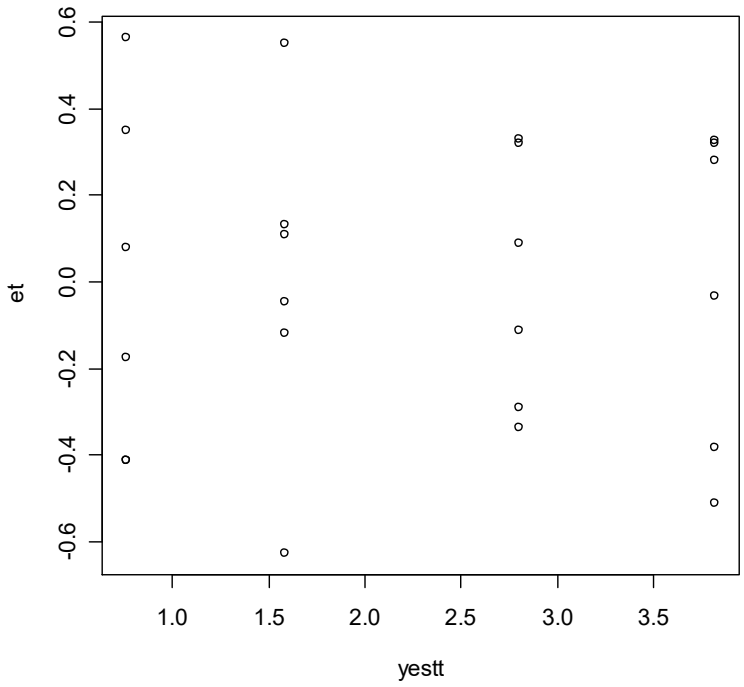



```
> mod<-lm(lsi~lyp)
> mod

Call:
lm(formula = lsi ~ lyp)

Coefficients:
(Intercept)          lyp
    -0.2781       0.4465
```

se puede usar la transformación raíz cuadrada

```
> yt<-y^0.5
```

```
> mod2<-lm(yt~metodo)
> anova(mod2)
Analysis of Variance Table

Response: yt
          Df Sum Sq Mean Sq F value    Pr(>F)
metodo     3 32.684  10.895  81.049 2.296e-11 ***
Residuals 20  2.688   0.134
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> et<-residuals(mod2)
> yestt<-predict(mod2)
> plot(yestt,et)
```




```
> bartlett.test(yt~metodo)

        Bartlett test of homogeneity of variances

data:  yt by metodo
Bartlett's K-squared = 0.5247, df = 3, p-value = 0.9134
```

Método Analítico para encontrar λ (Transformación de Box y Cox)

El método de Box y Cox es la manera más popular para determinar la transformación que se aplicar a la variable respuesta. Este método está diseñado estrictamente para valores positivos de la respuesta y elige la transformación para encontrar el mejor ajuste de la

respuesta de los datos. El método transforma la respuesta $Y \to t_\lambda(Y)$ donde la familia de transformaciones indexada por λ es

$$t_\lambda(Y) = \begin{cases} \dfrac{Y^\lambda - 1}{\lambda}, & \text{si } \lambda \neq 0 \\ \ln(Y), & \text{si } \lambda = 0 \end{cases}$$

Para valores fijado de $Y > 0$, la $t_\lambda(Y)$ es continua en λ. Se elige λ usando el método de máxima verosimilitud. El perfil del log-verosimilitud asume normalidad de los errores es

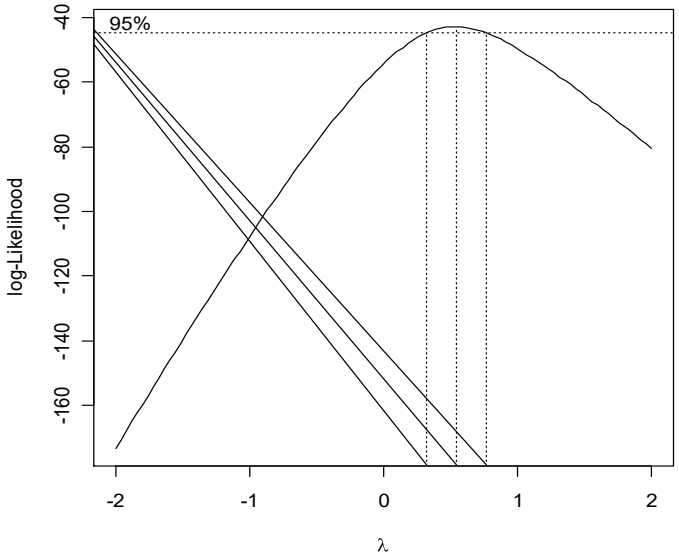
$$L(\lambda) = -\frac{n}{2}\ln(SCE_\lambda / n) + (\lambda - 1)\sum \ln(Y_i)$$

donde SCE_λ es la suma de cuadrado residual cuando $t_\lambda(Y)$ es la variable respuesta. Se puede calcular $\hat{\lambda}$ maximizando $L(\lambda)$, pero usualmente $L(\lambda)$ es maximizado sobre una malla de valores tales como $\{-2, -1, -1/2, 0, 1/2, 1, 2\}$. Esto asegura que se elija el valor de λ de manera que sea más fácilmente de interpretar. Por ejemplo si $\hat{\lambda} = 0.46$, podría usarse mejor la transformación de $\sqrt{Y}$, ya que es más fácil de interpretar.

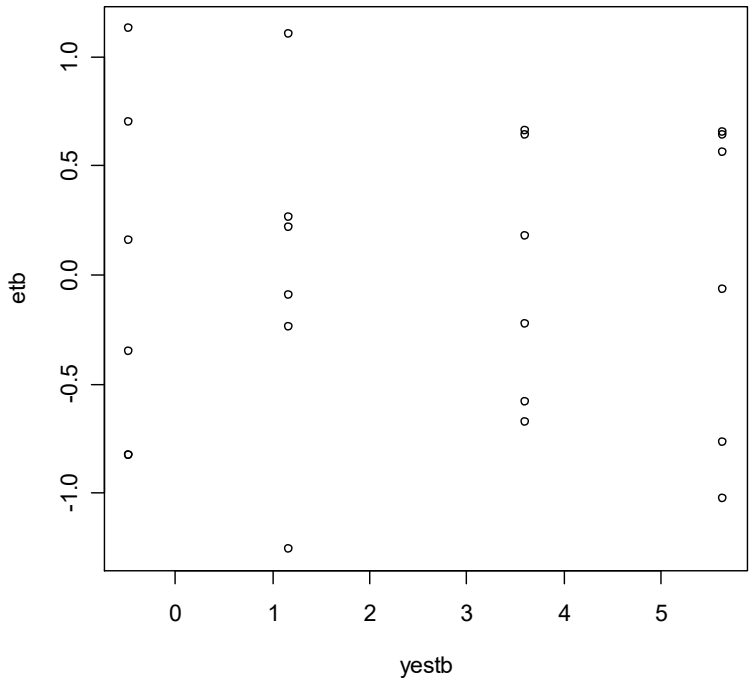
Nota Importante: Una vez transformado los datos todo el proceso de inferencia se realiza con los datos transformados.

Con los datos del ejemplo anterior se tiene:

```
> library(MASS)
> boxcox(y~metodo)
```



```
> ytb<-(y^0.5-1)/.5
> mod3<-lm(ytb~metodo)
> anova(mod3)
Analysis of Variance Table

Response: ytb
          Df  Sum Sq Mean Sq F value    Pr(>F)
metodo     3 130.737  43.579  81.049 2.296e-11 ***
Residuals 20  10.754   0.538
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> etb<-residuals(mod3)
> yestb<-predict(mod3)
> plot(yestb,etb)
```




```
> bartlett.test(ytb~metodo)

        Bartlett test of homogeneity of variances

data:  ytb by metodo
Bartlett's K-squared = 0.5247, df = 3, p-value = 0.9134
```