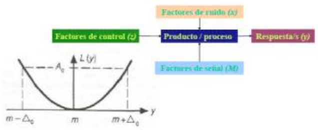
Transformaciones para estabilizar varianza y mejorar la normalidad

Transformación BOX-COX



Dr. Christian R. Encina Zelada

cencina@lamolina.edu.pe

- En la práctica, algunas variables de respuesta no siguen una distribución normal sino que se distribuyen, por ejemplo *Poisson*, *binomial* o *Gamma*, entre otras.
- Resulta que en estas distribuciones la media está relacionada con la desviación estándar (variabilidad) y, naturalmente, al cambiar la media de un tratamiento a otro, con ella cambia la variabilidad de la respuesta.
- También es cierto que al suponer normalidad y varianza constante, éstas no se tienen que cumplir de manera estricta, **dado que el procedimiento de ANOVA es robusto o admite desviaciones moderadas de dichos supuestos.**

- Existen al menos tres maneras de solucionar o minimizar el problema por falta de normalidad y de varianza heterogénea en los residuos:

1) utilizar **métodos de análisis no paramétricos**, que no requieren las suposiciones de normalidad y varianza constante;
2) hacer el análisis mediante **modelos lineales generalizados (GLM)**, en los que se ajusta un modelo lineal usando otras distribuciones diferentes a la normal, donde la varianza no tiene por qué ser constante, y
3) hacer el **análisis sobre la respuesta transformada** a una escala en la que los supuestos se cumplan.

Transformación apropiada	Tipo de transformación
$Y' = \text{sen}^{-1}(\sqrt{Y})$	Arcoseno, útil cuando la respuesta Y son proporciones (se distribuye binomial)
$Y' = \sqrt{Y}$	Raíz cuadrada, para los datos tipo Poisson
$Y' = \ln(Y)$ o $Y' = \log_{10}(Y)$	Transformación logaritmo
$Y' = Y^{-1/2}$	Recíproco de la raíz cuadrada
$Y' = Y^{-1}$	Recíproco

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 ($\ln(Y_{ij}) \sim$ Normal), entonces la transformación es logarítmica $\ln(Y_{ij})$.

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

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 transformados sea una constante, $\lambda = 1 - \alpha$. 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:

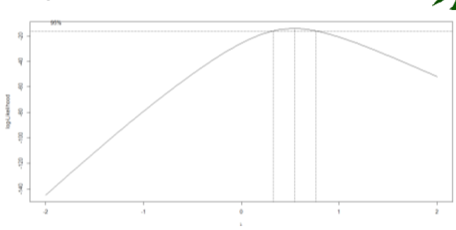
Relación entre σ_Y y μ	α	$\lambda = 1 - \alpha$	Transformación
$\sigma_Y \propto$ constante	0	1	Ninguna
$\sigma_Y \propto \mu^{1/2}$	1/2	1/2	Raíz cuadrada
$\sigma_Y \propto \mu$	1	0	Logarítmica
$\sigma_Y \propto \mu^{3/2}$	3/2	-1/2	Recíproca de la Raíz cuadrada
$\sigma_Y \propto \mu^2$	2	-1	Recíproca

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

- La transformación Box-Cox es una transformación de potencia que corrige la asimetría de una variable, diferentes varianzas o la no linealidad entre variables.
- En consecuencia, es muy útil transformar una variable y por tanto obtener una nueva variable que siga una distribución normal.



λ	Transformation
-2	$1/x^2$
-1	$1/x$
-0.5	$1/\sqrt{x}$
0	$\log(x)$
0.5	$\sqrt{x}$
1	x
2	x^2

$$\begin{cases} \frac{x^\lambda - 1}{\lambda} & \text{if} \quad \lambda \neq 0 \\ log(x) \text{ if} & \lambda = 0 \end{cases},$$

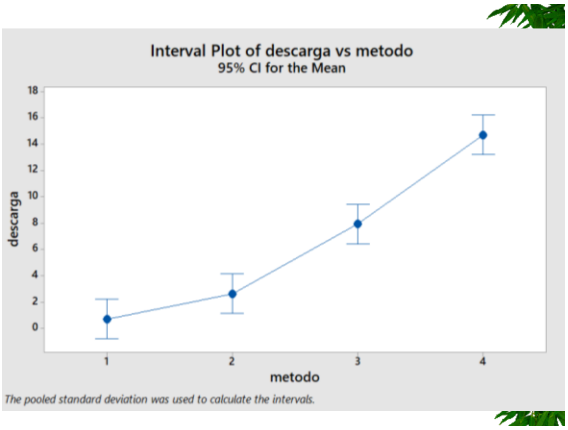
being y the changed variable and lambda (λ) the transformation parameter However, the following table describes the most typical transformations:

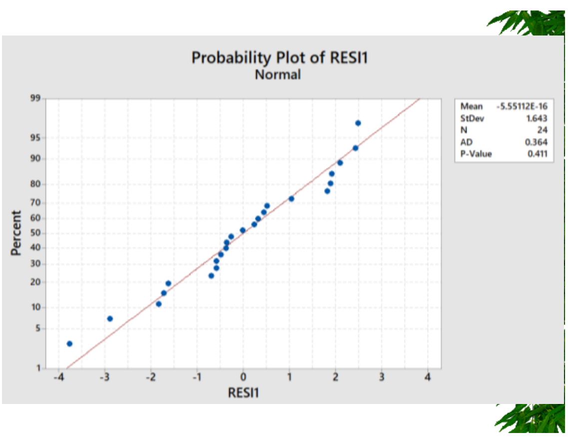
λ	Transformation
-2	1/x^2
-1	1/x
-0.5	1/sqrt(x)
0	log(x)
0.5	sqrt(x)
1	x
2	x^2

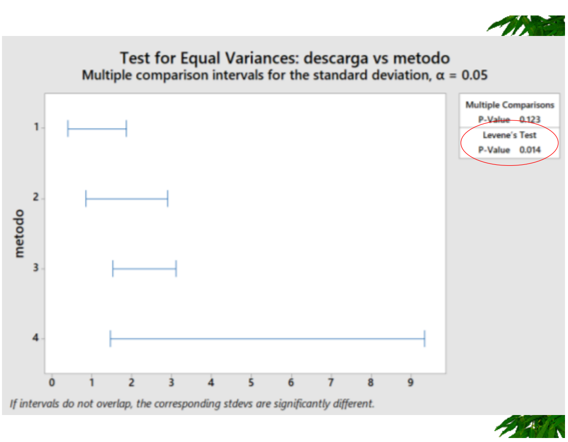
- Si el parámetro de transformación estimado está cerca de uno de los valores de la tabla anterior, en la práctica se recomienda elegir el valor de la tabla en lugar del valor exacto, ya que el valor de la tabla es más fácil de interpretar.
- Cuando usamos R, podemos utilizar la función “Box-Cox” del paquete MASS para estimar el parámetro de transformación mediante estimación de máxima verosimilitud.
- Esta función también nos dará el intervalo de confianza del 95% del parámetro. Los argumentos de la función son los siguientes:

Ejemplo: Un Ingeniero alimentario está interesado en determinar si cuatro métodos diferentes para el contenido de ácido ascórbico producen estimaciones equivalentes. Cada procedimiento se usa seis veces y los datos de ácido ascórbico (en mg/100 ml) se muestran en la siguiente tabla:

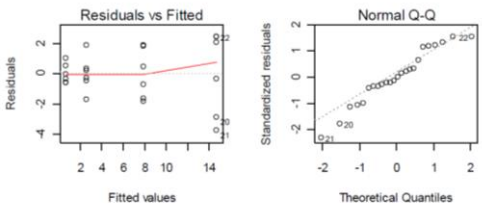
Método de Estimación	Observaciones					
1	0.34	0.12	1.23	0.70	1.75	0.12
2	0.91	2.94	2.14	2.36	2.86	4.55
3	6.31	8.37	9.75	6.09	9.82	7.24
4	17.15	11.82	10.95	17.20	14.35	16.82

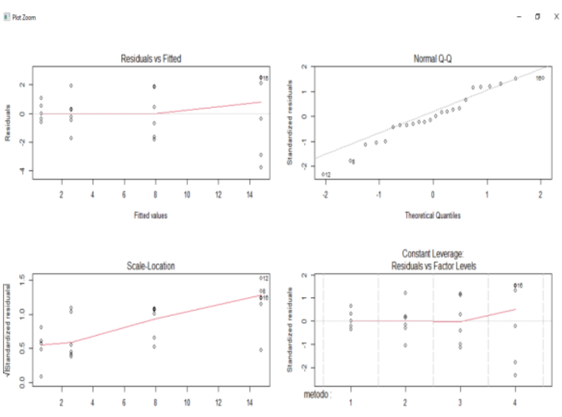

Interval Plot of descarga vs metodo
95% CI for the Mean
descarga
metodo
The pooled standard deviation was used to calculate the intervals.


Probability Plot of RESI1
Normal
Percent
RESI1
Mean -5.55112E-16
StDev 1.643
N 24
AD 0.364
P-Value 0.411


Test for Equal Variances: descarga vs metodo
Multiple comparison intervals for the standard deviation, α = 0.05
metodo
Multiple Comparisons
P-Value 0.123
Levene's Test
P-Value 0.014
If intervals do not overlap, the corresponding stdevs are significantly different.

```
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
```



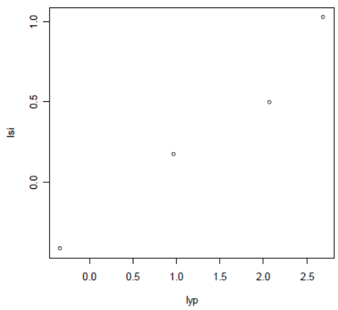


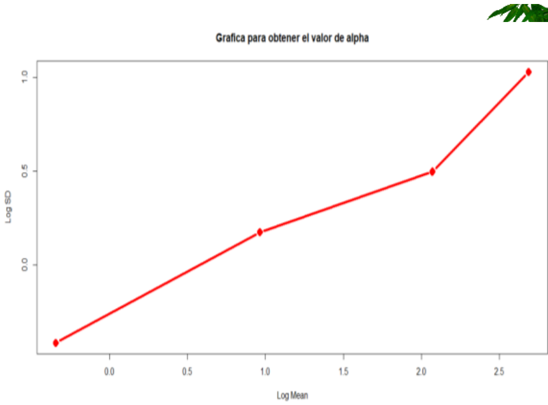

```
> ncvTest(mod1)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 9.604614    Df = 1     p = 0.001940891

> leveneTest(mod1)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value  Pr(>F)
group  3  4.5684 0.01357 *
      20
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
1
```

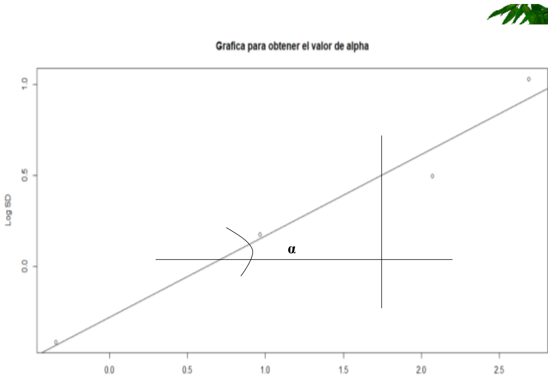
```
Bartlett's K-squared = 8.9958, df = 3, p-value = 0.02935
```

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)
```





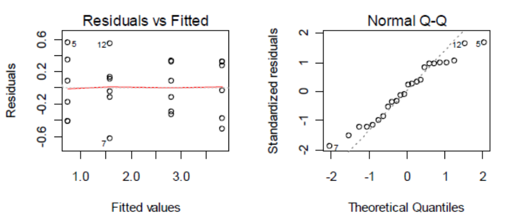


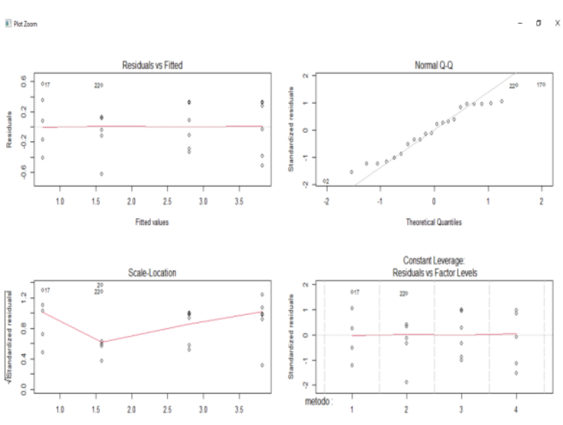
```
> mod<-lm(lsi~lyp)
> mod

Call:
lm(formula = lsi ~ lyp)

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

"α"

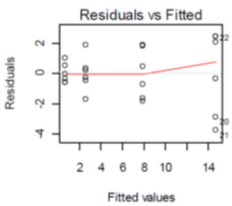
```
se puede usar la transformación raíz cuadrada ya que
λ=1-α=1-0.4465=0.5535

> yt<-y^0.5
> mod2<-lm(yt~metodo)
> anova(mod2)
Analysis of Variance Table
```

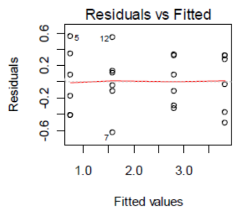
```
Response: yt
          Df Sum Sq Mean Sq F value    Pr(>F)
metodo     3 32.684  10.895  81.049 2.296e-11 ***
Residuals 20  2.688   0.134
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> par(mfrow=c(2,2))
> plot(mod2)
```





VARIANZAS

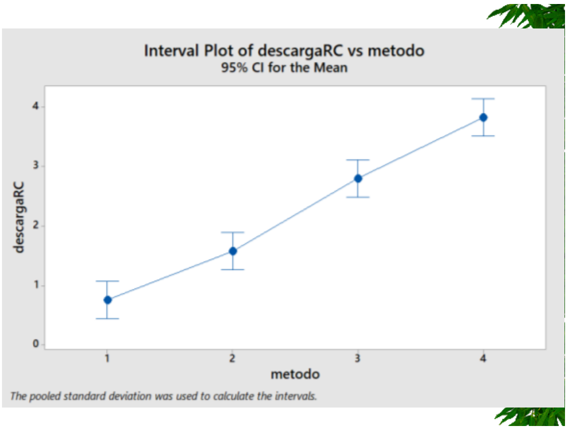


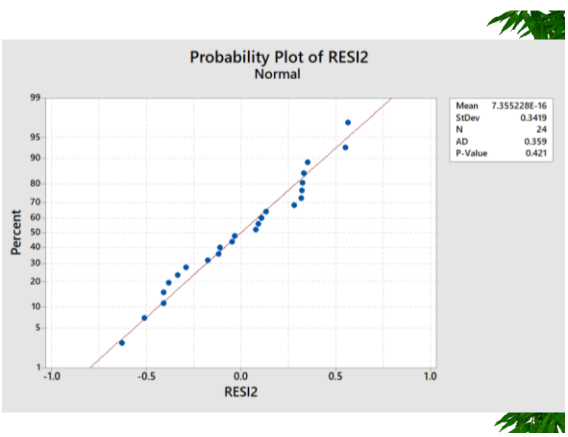


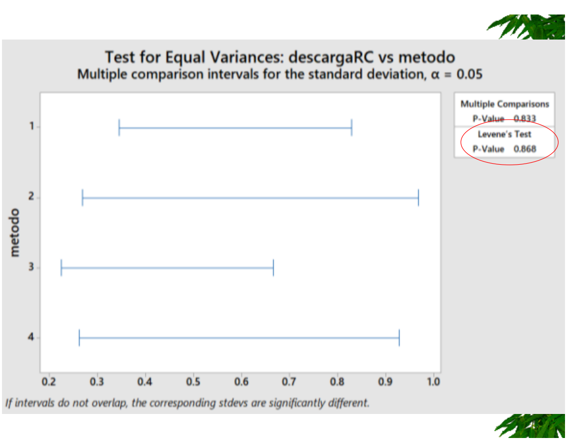
Antes **Después**

```
> 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

> ncvTest(mod2)
Non-constant Variance Score Test
```

```
Variance formula: ~ fitted.values
Chisquare = 0.1582841     Df = 1     p = 0.6907412

> ri<-rstandard(mod2)
> shapiro.test(ri)

        Shapiro-Wilk normality test

data:  ri
W = 0.9588, p-value = 0.4141

> 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
```


Interval Plot of descargaRC vs metodo
95% CI for the Mean
descargaRC
0
1
2
3
4
metodo
1
2
3
4
The pooled standard deviation was used to calculate the intervals.


Probability Plot of RESI2
Normal
Percent
1
5
10
20
30
40
50
60
70
80
90
95
99
RESI2
-1.0
-0.5
0.0
0.5
1.0
Mean 7.355228E-16
StDev 0.3419
N 24
AD 0.359
P-Value 0.421


Test for Equal Variances: descargaRC vs metodo
Multiple comparison intervals for the standard deviation, α = 0.05
metodo
1
2
3
4
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Multiple Comparisons
P-Value 0.833
Levene's Test
P-Value 0.868
If intervals do not overlap, the corresponding stdevs are significantly different.

- Cabe aclarar que las transformaciones para estabilizar la varianza no eliminan el efecto de dispersión que de por sí existe.

- Sólo permiten analizar mejor el efecto sobre la media.

- La familia de diseños factoriales completos 2^k (k factores con dos niveles de prueba cada uno), que es una de las familias de diseños de mayor impacto en la industria y en la investigación, debido a su eficacia y versatilidad.
- Los *factoriales* 2^k *completos* son útiles principalmente cuando el número de factores a estudiar está entre dos y cinco ($2 \leq k \leq 5$), rango en el cual su tamaño se encuentra entre cuatro y 32 tratamientos; esta cantidad es manejable en muchas situaciones experimentales.