BIOMETRÍA II

CLASE 10

DISEÑOS ANIDADOS

Adriana Pérez
Depto de Ecología, Genética y Evolución
FECN, UBA

Efecto del pastoreo sobre el banco de semillas de un pastizal

- Se seleccionaron diez potreros de 1ha cada uno. Cada potrero fue asignado al azar a uno de dos tratamientos: régimen de pastoreo por ganado bovino o régimen de clausura de ganado durante 5 años, en un diseño balanceado
- En cada parcela se eligieron 10 puntos al azar y en cada uno se extrajo con un barreno de 5 cm de diámetro por 5 cm de altura una muestra de suelo y se determinó en cada muestra la biomasa de semillas (en gramos/m^2)

Experimento o estudio observacional?

VR:

Tipo? Potencial distribución de probabilidades?

VE:

Tipo? De efectos fijos o aleatorios?

Datos (gramos semillas/m²)

	tratamiento	potrero	biomasa
1	pastoreo	1	8.2
2	pastoreo	1	8.8
3	pastoreo	1	9.5
4	pastoreo	1	12.7
5	pastoreo	1	15.2
6	pastoreo	1	13.0
7	pastoreo	1	8.5
8	pastoreo	1	6.7
9	pastoreo	1	9.9
10	pastoreo	1	8.5
11	pastoreo	2	4.9
12	pastoreo	2	10.5
13	pastoreo	2	7.5
14	pastoreo	2	9.0
15	pastoreo	2	6.4
16	pastoreo	2	8.5
17	pastoreo	2	4.9
18	pastoreo	2	3.7
19	pastoreo	2	6.6
20	pastoreo	2	7.5

Showing 1 to 20 of 100 entries

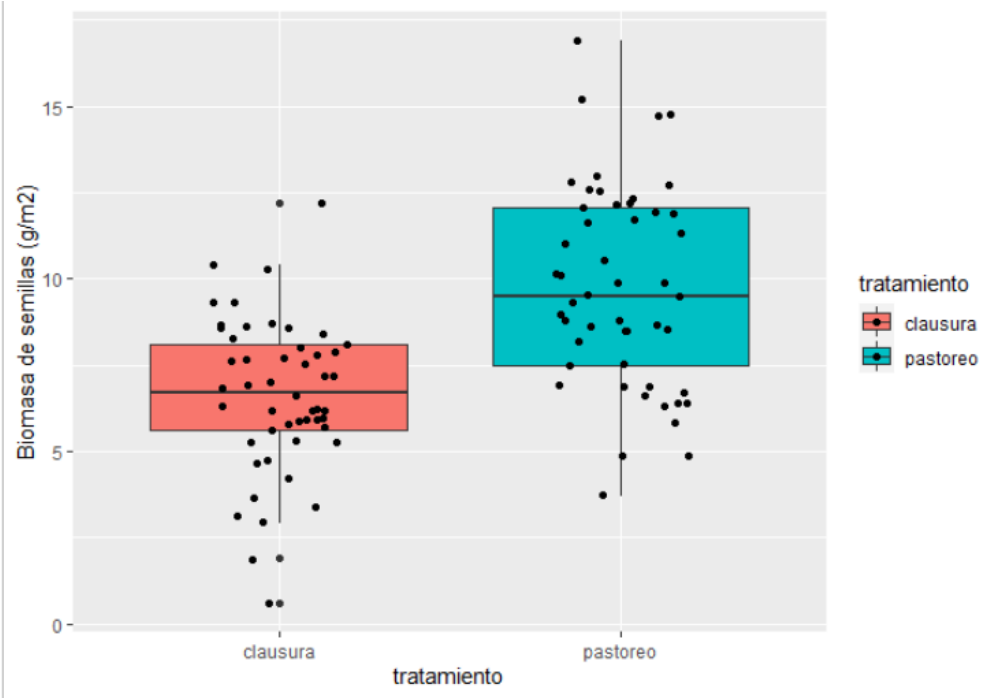
	tratamiento	potrero	biomasa
82	clausura	9	9.3
83	clausura	9	7.9
84	clausura	9	10.4
85	clausura	9	6.2
86	clausura	9	5.3
87	clausura	9	6.3
88	clausura	9	4.7
89	clausura	9	8.4
90	clausura	9	6.6
91	clausura	10	8.6
92	clausura	10	12.2
93	clausura	10	9.3
94	clausura	10	8.6
95	clausura	10	8.1
96	clausura	10	7.2
97	clausura	10	10.3
98	clausura	10	8.0
99	clausura	10	7.7
100	clausura	10	7.8

Showing 81 to 100 of 100 entries

semillas.csv

Opción 1
Ignorando los potreros

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$
i= 1,2
j=1 a 50
$$\varepsilon_{ij} \sim N(0, \sigma^2)$$



tratamiento	n	media	DE	EE
<fct>	<int>	<dbl>	<dbl>	<dbl>
clausura	50	6.64	2.21	0.313
pastoreo	50	9.74	2.93	0.414

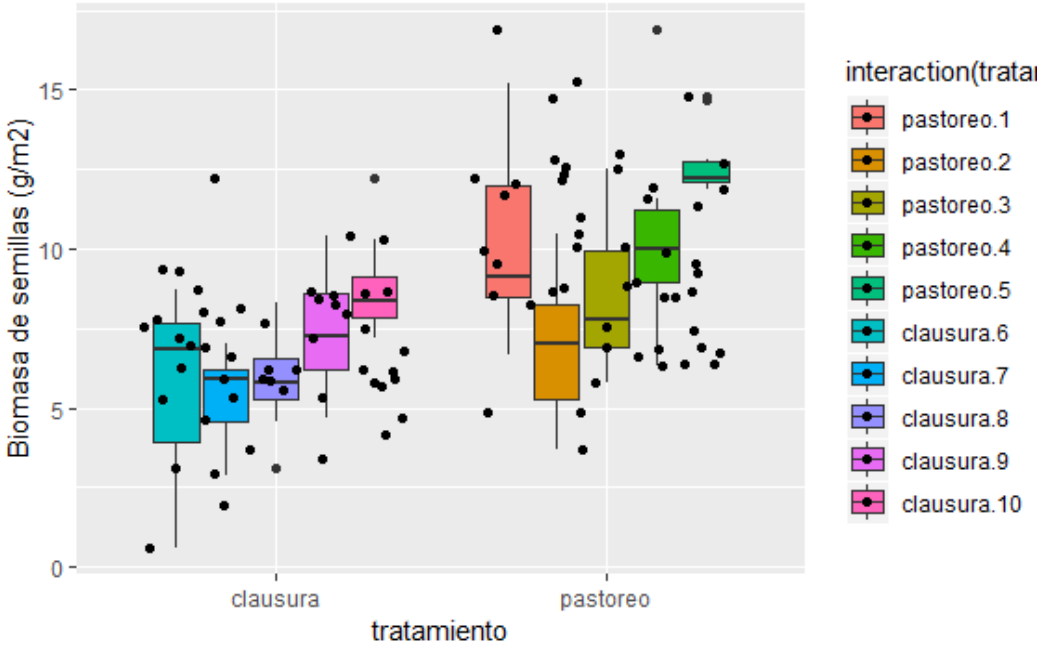
Seudoreplicación!
Los EE están subestimados, *p* menores a lo correcto, mayor probabilidad de error tipo I

```
m1<-lm(biomasa~tratamiento, semillas)
anova(m1)
Analysis of Variance Table
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
tratamiento	1	240.3	240.25	35.69	3.73e-08 ***
Residuals	98	659.7	6.73		

Opción 2
Promediando la VR por potrero

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$
$i = 1,2$
$j = 1 \text{ a } 5$
$\varepsilon_{ij} \sim N(0, \sigma^2)$



	tratamiento	potrero	biomasa
1	pastoreo	1	10.10
2	pastoreo	2	6.95
3	pastoreo	3	8.54
4	pastoreo	4	10.38
5	pastoreo	5	12.75
6	clausura	6	5.78
7	clausura	7	5.47
8	clausura	8	5.81
9	clausura	9	7.38
10	clausura	10	8.78

tratamiento	n	media	DE	EE
<fct>	<int>	<dbl>	<dbl>	<dbl>
clausura	5	6.64	1.41	0.629
pastoreo	5	9.74	2.17	0.970

```
m2<-lm(biomasa~tratamiento, medias.potrero)
anova(m2)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
tratamiento	1	24.02	24.025	7.186	0.0279	*
Residuals	8	26.75	3.343			

✔ Pero se pierde información

Opción 3. Modelo condicional (mixto) Incorporando la variable potrero al modelo

$Y_{ijk} = \mu + \alpha_i + Bj + \varepsilon_{ijk}$

i= 1,2

j=1 a 5

K= 1 a 10

$\varepsilon_{ijk} \sim N(0, \sigma^2)$

$B_j \sim N(0, \sigma^2_{potreros})$

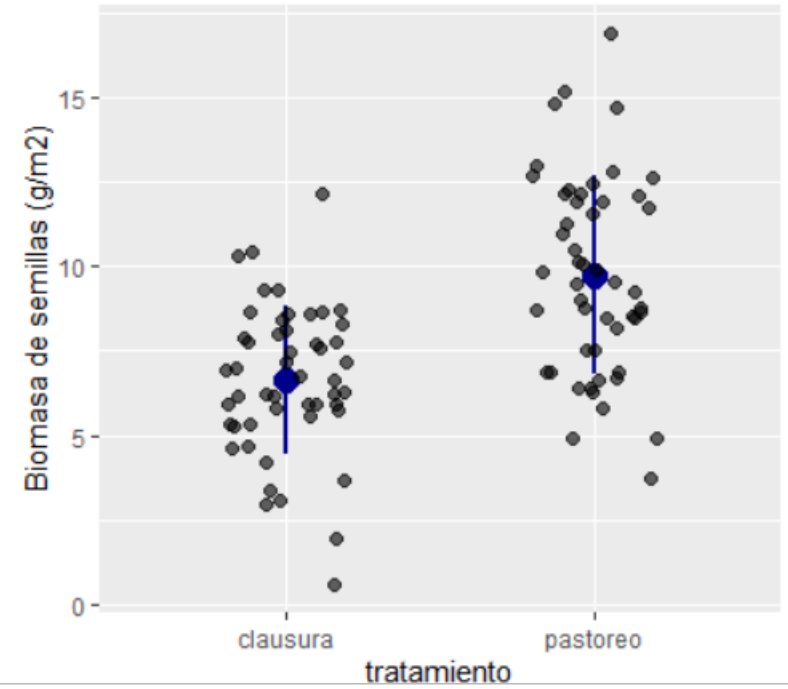
ε_{ij}, B_j $indep$

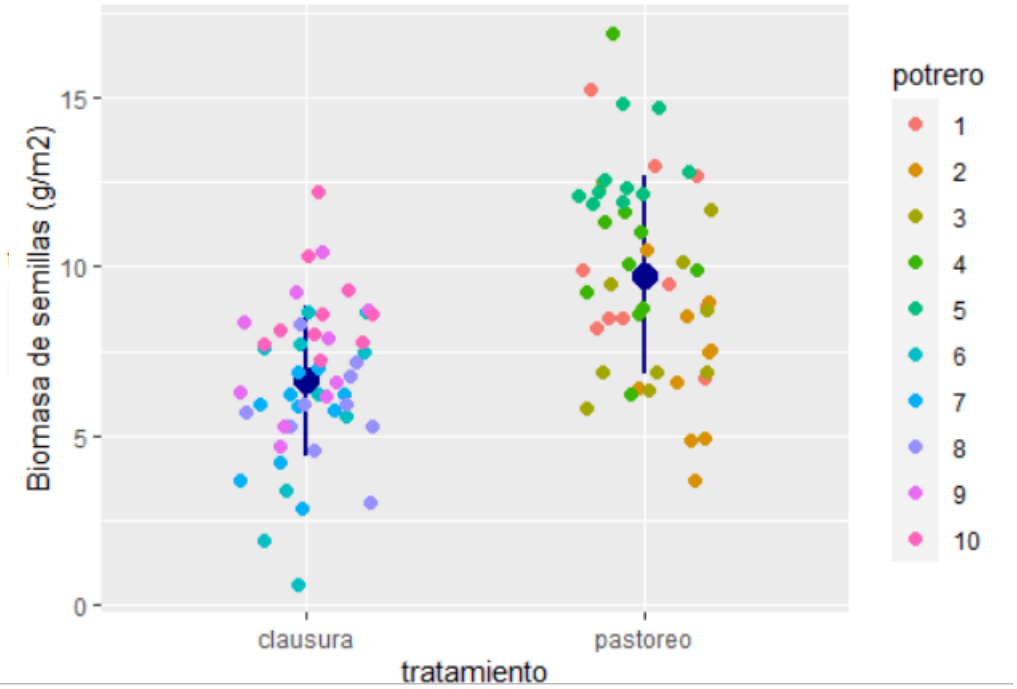
las observaciones son condicionalmente independientes, pero marginalmente estarán correlacionadas debido al efecto aleatorio.

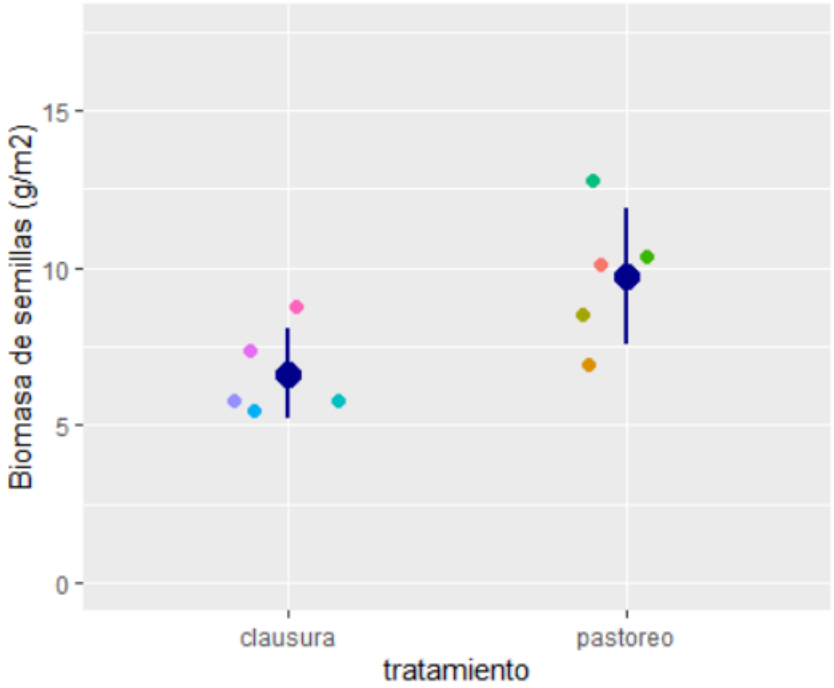
```
library(lme4)
m3<- lmer(biomasa ~ tratamiento + (1|potrero), semillas)
```

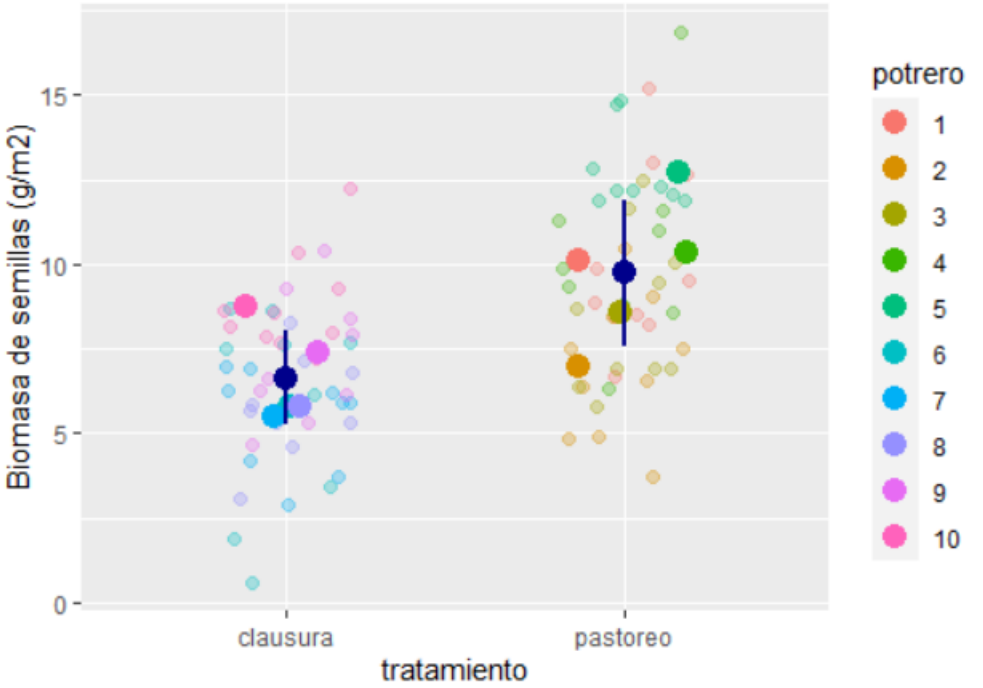
```
> anova(m3)
Type III Analysis of Variance Table with Satterthwaite's method
            Sum Sq Mean Sq NumDF DenDF F value Pr(>F)
tratamiento 31.314  31.314     1     8  7.1856 0.0279 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Parámetros? Efectos aleatorios?


Biomasa de semillas (g/m2)
15
10
5
0
clausura
pastoreo
tratamiento


Biomasa de semillas (g/m2)
15
10
5
0
clausura
pastoreo
tratamiento
potrero
1
2
3
4
5
6
7
8
9
10


Biomasa de semillas (g/m2)
15
10
5
0
clausura
pastoreo
tratamiento


Biomasa de semillas (g/m2)
15
10
5
0
clausura
pastoreo
tratamiento
potrero
1
2
3
4
5
6
7
8
9
10

Opción 3. Modelo condicional (mixto) Incorporando la variable potrero al modelo


```
library(lme4)
m3<- lmer(biomasa ~ tratamiento + (1|potrero), semillas)
Summary(m3)
```

```
Linear mixed model fit by REML t-tests use Satterthwaite approximations to
  degrees of freedom [lmerMod]
Formula: biomasa ~ tratamiento + (1 | potrero)
   Data: semillas

REML criterion at convergence: 446.5

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.5353 -0.6256 -0.0507  0.6122  3.1630

Random effects:
 Groups   Name        Variance Std.Dev.
 potrero  (Intercept) 2.908    1.705
 Residual             4.358    2.088
Number of obs: 100, groups:  potrero, 10

Fixed effects:
                  Estimate Std. Error     df t value Pr(>|t|)
(Intercept)         6.6440     0.8177 8.0000   8.125 3.91e-05 ***
tratamientopastoreo 3.1000     1.1565 8.0000   2.681   0.0279 *
---
```

Mismos resultados usando la función lme del paquete nlme

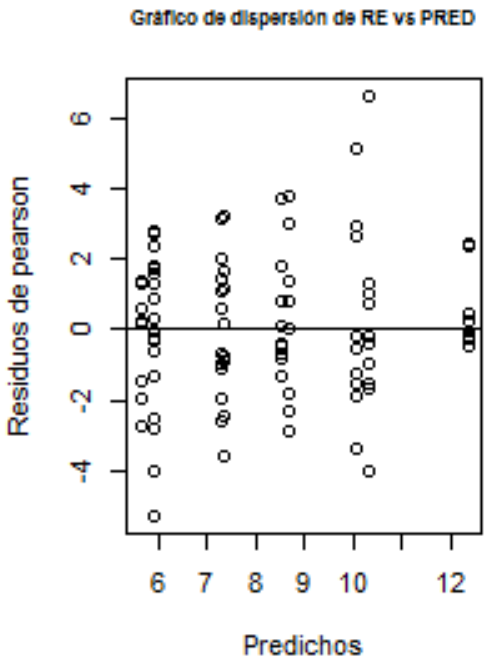
Opción 4.
Modelo marginal

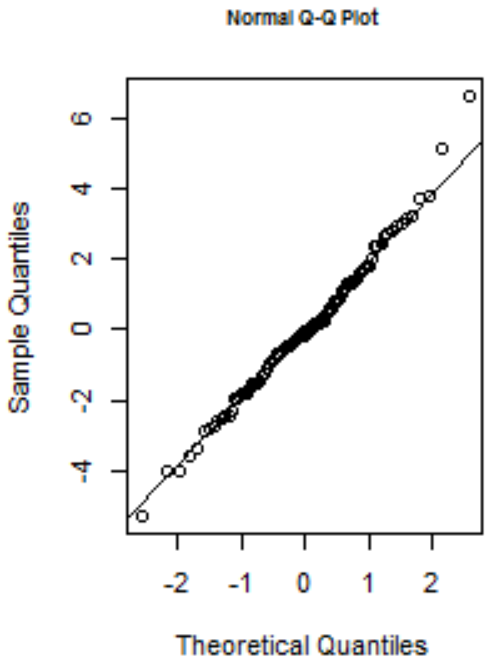

```
m4 <- gls(biomasa ~ tratamiento, correlation=corCompSymm(form = ~
1 | potrero), data = semillas)

Generalized least squares fit by REML
  Model: biomasa ~ tratamiento
  Data: semillas
       AIC      BIC    logLik
  454.4912 464.8311 -223.2456

Correlation Structure: Compound symmetry
 Formula: ~1 | potrero
 Parameter estimate(s):
      Rho
0.4002022

Coefficients:
                    Value Std.Error  t-value p-value
(Intercept)         6.644 0.8177383 8.124848  0.0000
tratamientopastoreo 3.100 1.1564566 2.680602  0.0086
```

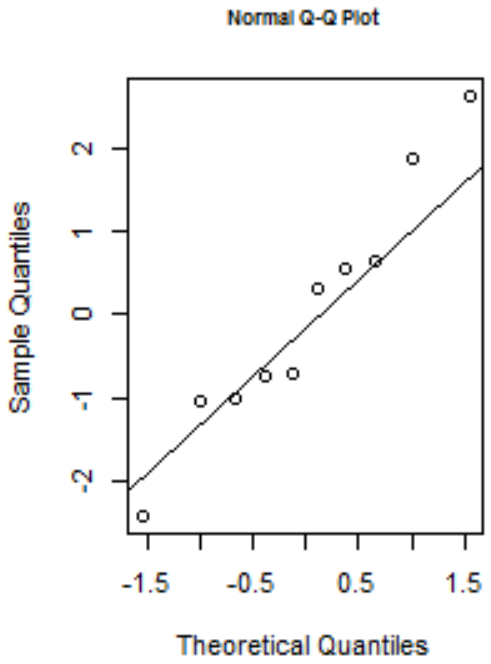
Mismos resultados para la parte fija que lmer y lme

Parámetros? Efectos aleatorios?

Supuestos








```
Shapiro-Wilk normality test data:
e W = 0.98998, p-value = 0.6632

Shapiro-Wilk normality test data:
alfai W = 0.9601, p-value = 0.787
```

Parte fija Significación?

Problemas:

- No hay consenso sobre los GL
- Las distribuciones de los estadísticos son asintóticas

- Prueba de Wald

```
Fixed effects:
                    Estimate Std. Error     df t value Pr(>|t|)
(Intercept)           6.6440     0.8177 8.0000   8.125 3.91e-05 ***
tratamientopastoreo   3.1000     1.1565 8.0000   2.681   0.0279 *
---
```

Comparar con opción 2

- Prueba de cociente de verosimilitud (Likelihood ratio test LRT):

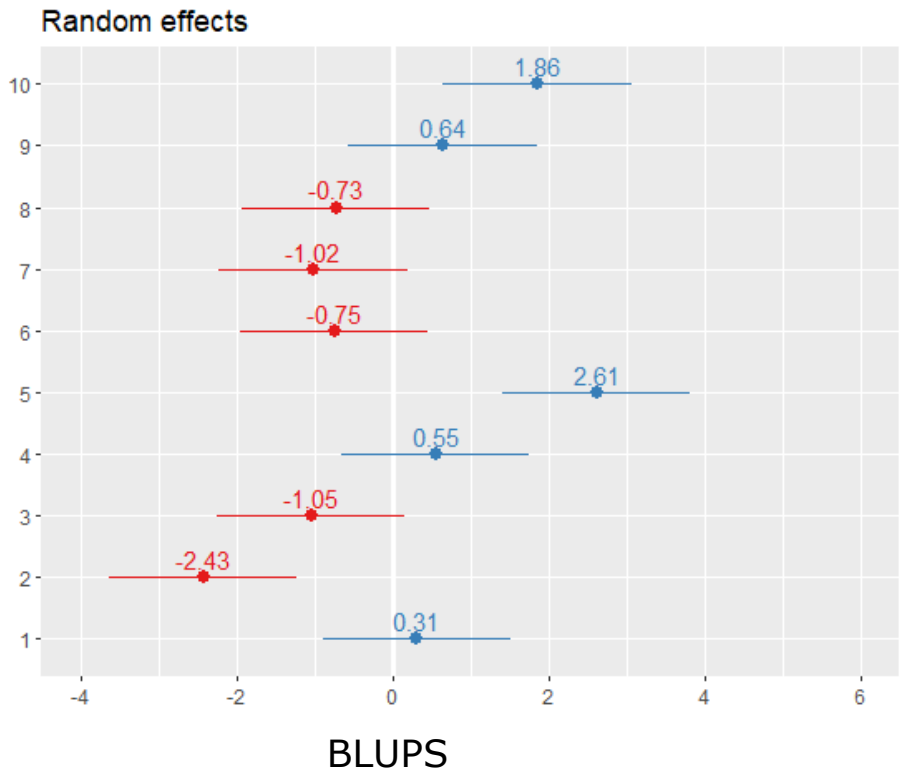
  Permite comparar modelos anidados con distinta estructura fija. Ojo, la estimación debe ser por máxima verosimilitud (no por REML). Se puede usar drop1 o anova

```
> m0<- lmer(biomasa ~  (1|potrero), semillas)
> anova(m0,m3)
refitting model(s) with ML (instead of REML)
Data: semillas
Models:
m0: biomasa ~ (1 | potrero)
m3: biomasa ~ tratamiento + (1 | potrero)
   Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
m0  3 461.54 469.36 -227.77   455.54
m3  4 457.13 467.55 -224.56   449.13 6.4091      1    0.01135 *
```

Parte aleatoria
Efectos aleatorios

$$BLUP_j = BLUE_j \left(\frac{\sigma^2_{potreros}}{\sigma^2_{potreros} + \sigma^2 / n_i} \right)$$

```
$potrero
   (Intercept)
1    0.3095992
2   -2.4298321
3   -1.0470715
4    0.5531042
5    2.6142002
6   -0.7513869
7   -1.0209817
8   -0.7252971
9    0.6400703
10   1.8575954
```



Parte aleatoria
Componentes de varianza

```
Random effects:
 Groups   Name        Variance Std.Dev.
 potrero  (Intercept) 2.908    1.705
 Residual             4.358    2.088
Number of obs: 100, groups:  potrero, 10
```

$$\hat{\sigma}^2_{Y_{ij}} = 2,908 + 4,358 = 7,266$$

$$\hat{\sigma}_{Y_{ij}} = 2,7 g / m^2$$

$$CCI = \frac{\sigma^2_{potrero}}{\sigma^2_{potrero} + \sigma^2} = 0,4$$

El 40% de la variación en la biomasa de semillas está aportada por la variación entre potreros sometidos a un determinado tratamiento; el 60% restante está aportado por la variación entre muestras dentro de un mismo potrero

El coeficiente de correlación intraclase CCI mide la correlación entre puntos de un mismo lote; cuanto más alta sea indica que las mediciones dentro de un mismo lote son muy similares y por lo tanto la variación viene dada por los potreros

Parte aleatoria Significación?

- Según algunos autores, si el efecto aleatorio está dado por diseño, debería permanecer en el modelo
- Puede usarse la prueba de cociente de verosimilitud, tanto con ML o REML, ya que estamos comparando modelos con la misma parte fija. La prueba es conservativa.

```
> ranova(m3)
ANOVA-like table for random-effects: Single term deletions

Model:
biomasa ~ tratamiento + (1 | potrero)
              npar  logLik    AIC    LRT Df Pr(>Chisq)
<none>           4 -223.25 454.49
(1 | potrero)    3 -236.40 478.80 26.311  1  2.906e-07 ***
---
```

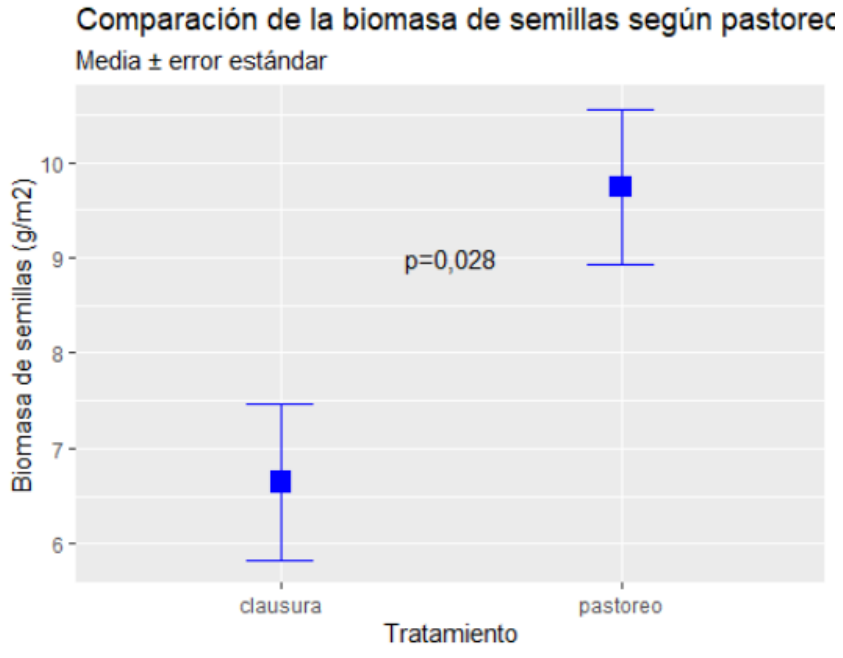
- Al probar si una o más varianzas son cero estamos en la frontera del espacio de parámetros (ya que el mínimo de una varianza es cero), y por lo tanto la distribución asintótica del estadístico de la prueba es aproximada
- Se pueden construir intervalos de confianza para los componentes de varianza del método usando el método de verosimilitud perfilada (profile máximum likelihood)

```
confint(mod.mix4, level = 0.95, method = c("profile"))
                           2.5 %   97.5 %
.sig01                 0.8935747 2.650294
.sigma                 1.8160876 2.434040
```

Presentación de resultados


```
confint(emmeans(m3, pairwise ~ tratamiento))
```

```
$emmeans
 tratamiento emmean    SE df lower.CL upper.CL
 clausura      6.64 0.818  8     4.76     8.53
 pastoreo      9.74 0.818  8     7.86    11.63

Degrees-of-freedom method: kenward-roger
Confidence level used: 0.95

$contrasts
 contrast            estimate   SE df lower.CL upper.CL
 clausura - pastoreo     -3.1 1.16  8    -5.77   -0.433

Degrees-of-freedom method: kenward-roger
Confidence level used: 0.95
```



¿Para qué sirvió anidar?

- Obtener submuestras es usualmente menos costoso que incrementar la cantidad de ue
- Pero ojo: el submuestreo no incrementa el número de réplicas. El número de réplicas sigue siendo el número de UE a las que se le aplicó el tratamiento en forma aleatoria, y no el número de observaciones por tratamiento
- Por lo tanto el aumento en la cantidad de submuestras no incrementa directamente la potencia de la prueba (ésta depende de la cantidad de réplicas)
- Sin embargo, si existe mucha variación a pequeña escala (elevado σ^2), si se aumenta la cantidad de submuestras la estimación de la variación entre UE será más precisa, lo que indirectamente aumentará la potencia de la prueba
- Además, si existe desbalanceo en las submuestras, este análisis provee estimaciones que lo toman en cuenta
- Los BLUP, es decir los efectos aleatorios, se encogen (se parecen más a la media general) si:
  - el componente de varianza para el término en cuestión es pequeño
  - la varianza residual es grande
  - el número de repeticiones del nivel de factor considerado es pequeño
- Las submuestras son en muchos casos réplicas técnicas

Estimación por MV restringida vs MV


```
m2 <- lmer(biomasa ~ tratamiento
+ (1|potrero), data = semillas,
REML=TRUE)
```

```
Random effects:
 Groups   Name        Variance Std.Dev.
 potrero  (Intercept) 2.908    1.705
 Residual             4.358    2.088
Number of obs: 100, groups:  potrero, 10

Fixed effects:
                   Estimate Std. Error
(Intercept)          6.6440     0.8177
tratamientopastoreo  3.1000     1.1565
```

Es el método por defecto

```
m2ML <- lmer(biomasa ~ tratamiento
+ (1 | potrero), data = semillas,
REML=FALSE)
```

```
Random effects:
 Groups   Name        Variance Std.Dev.
 potrero  (Intercept) 2.239    1.496
 Residual             4.358    2.088
Number of obs: 100, groups:  potrero, 10

Fixed effects:
                   Estimate Std. Error t
(Intercept)          6.6440     0.7314
tratamientopastoreo  3.1000     1.0344
```

Si se utiliza estimación por MV:
Las varianzas y EE están subestimados, pero no los estimadores de los coeficientes para VE de efectos fijos
Solo es indicada para comparar modelos

Factores cruzados vs anidados

- Dos factores (VE cualitativas) están cruzados cuando cada nivel de un factor está observado en todos los niveles del otro (y viceversa). Corresponde a un diseño factorial. No hay jerarquía
- El factor B está anidado en A cuando cada nivel del factor B está observado en un solo nivel de A (hay jerarquía). Como cada nivel de B no se cruza con cada nivel de A, no es posible que exista interacción entre A y B
- Bloques? Potreros?
- Para que R detecte que los potreros están anidados en los tratamientos y no cruzados, se los debe identificar unívocamente:
  - Potreros 1 a 10 => anidados en tratamiento
  - Potreros 1 a 5 en Control y 1 a 5 en Clausura => cruzados con tratamiento

$$Y_{ij} = \mu + \alpha_i + B_j + \varepsilon_{ij}$$
$$\varepsilon_{ij} \approx NID(0, \sigma^2)$$
$$B_j \approx NID(0, \sigma^2{}_{potreros})$$

Variaciones en rasgos del cedro amargo

- Se llevó a cabo un estudio en el NOA a fin de caracterizar la variabilidad fenotípica en el cedro americano (*Cedrela odorata*) , una especie vulnerable.
- Se estudiaron 7 poblaciones elegidas al azar en el área de estudio. De cada población se eligieron entre 12 y 20 familias y de cada familia se estudiaron al menos dos ejemplares.
- Se registró el largo de cada ejemplar

Experimento o estudio observacional?

VR:

Tipo? Potencial distribución de probabilidades?

VE:

Tipo? De efectos fijos o aleatorios?

Agrupamiento?

cedro.csv

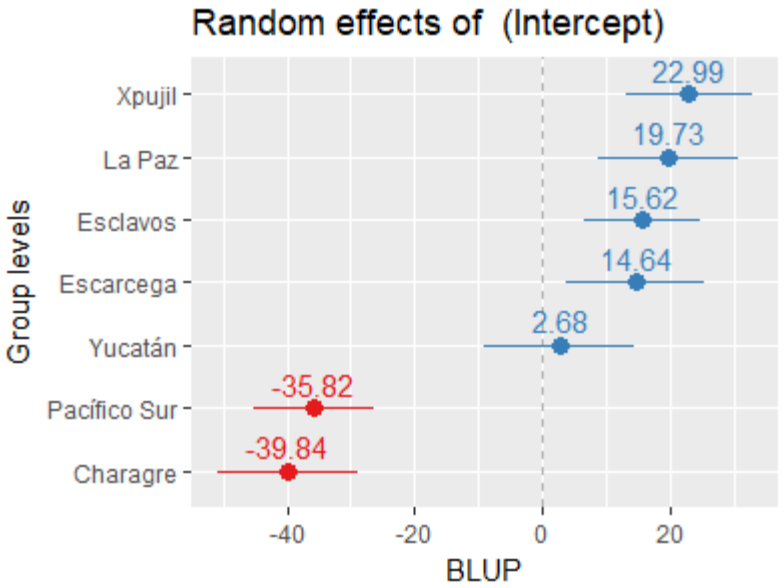
Diseño totalmente anidado

Totalmente anidado: Factor A (aleatorio), Factor B anidado en A, Factor C anidado en B

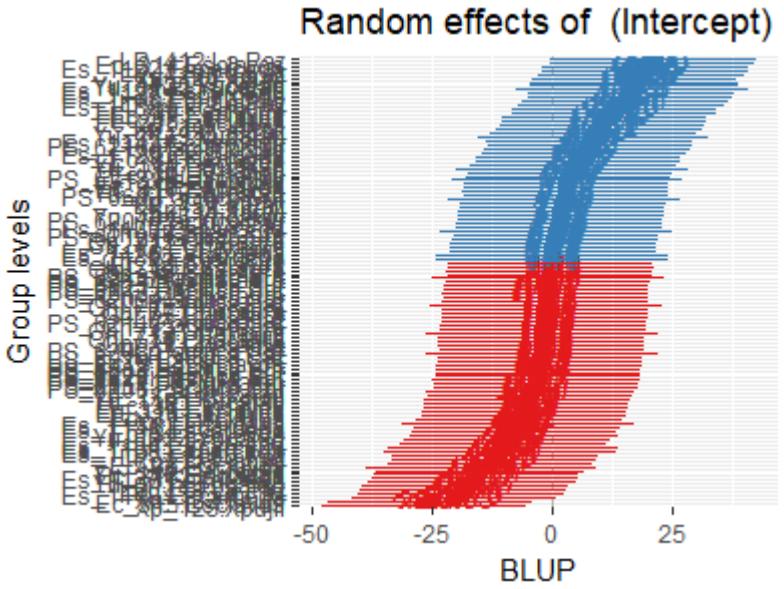
```
lmer(largo ~ 1 + (1 | poblacion/familia), BD)
lmer(largo ~ 1 + (1 | poblacion)+ (1 | familia), BD)
lme(largo ~ 1, random = ~ 1|poblacion/familia, BD)
```

```
> BD
       poblacion familia largo
1       Charagre   Ch_71   6.0
2       Charagre   Ch_71   6.0
3       Charagre  Ch_710   6.0
4       Charagre  Ch_710  13.0
5       Charagre  Ch_711  14.0
6       Charagre  Ch_711   8.0
7       Charagre  Ch_712  12.5
8       Charagre  Ch_712  10.0
9       Charagre  Ch_713   6.5
10      Charagre  Ch_713   6.0
```

```
> summary(m4)
Linear mixed model fit by REML ['lmerMod']
Formula: largo ~ 1 + (1 | poblacion/familia)
   Data: BD

REML criterion at convergence: 2008.5

Scaled residuals:
     Min       1Q   Median       3Q      Max
-2.24033 -0.42502 -0.05879  0.55051  2.43795

Random effects:
 Groups             Name        Variance Std.De
 familia:poblacion (Intercept) 219.0    14.80
 poblacion         (Intercept) 737.5    27.16
 Residual                      463.7    21.53
Number of obs: 214, groups:  familia:poblacion, 115; poblacion, 7

Fixed effects:
            Estimate Std. Error t value
(Intercept)    49.85      10.47   4.762
```



¿Qué miden?
¿Cuánto aportan?
¿Cuáles son sus unidades?



Complicando el modelo

- ¿Y si de cada ejemplar se eligieron 10 semillas al azar y se registró el peso de cada una?

- ¿Y si de cada ejemplar se registró el pH del suelo e interesa saber si el largo del ejemplar se asocia con el pH?

- ¿Y si se sospecha que el "efecto" del pH sobre el largo del ejemplar cambia entre poblaciones?

Complicando el modelo

- ¿Y si de cada ejemplar se eligieron 10 semillas al azar y se registró el peso de cada una?

```
lmer(peso ~ 1 + (1 | poblacion/familia/ejemplar))
lme(peso ~ 1, random = ~ 1|poblacion/familia/ejemplar)
```

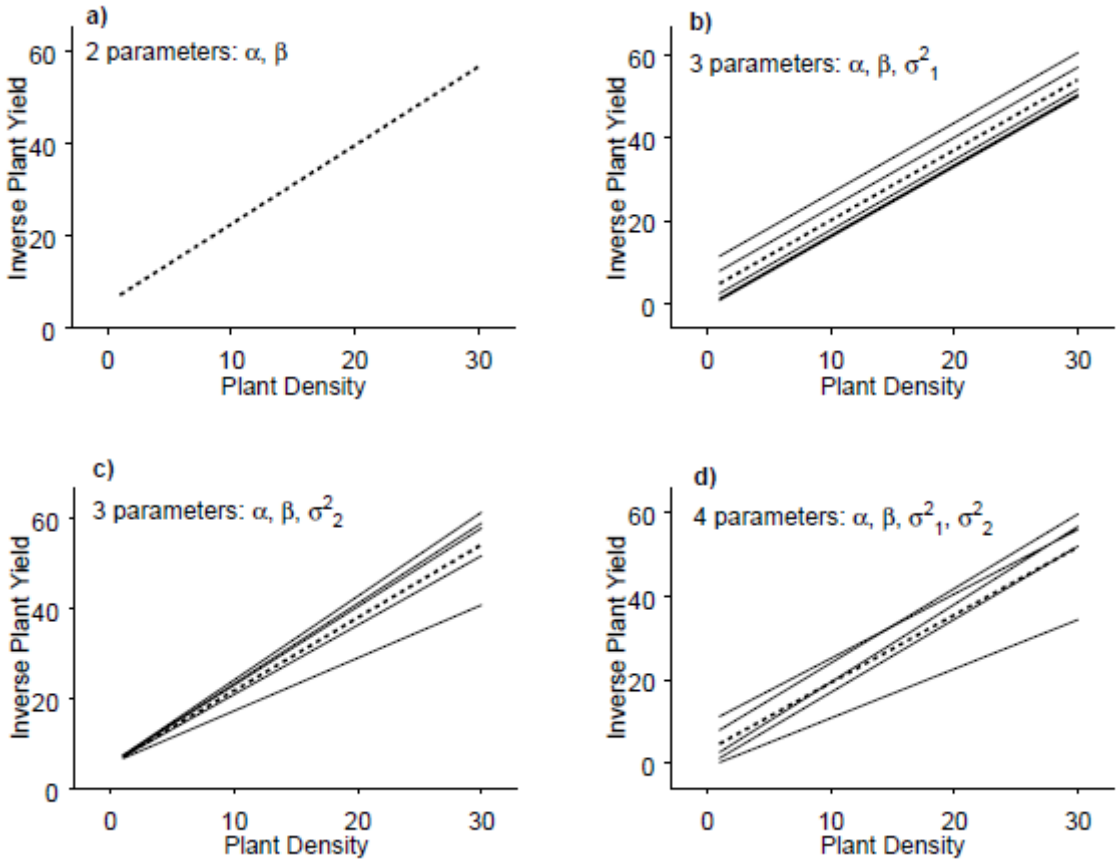
- ¿Y si de cada ejemplar se registró el pH del suelo e interesa saber si el largo del ejemplar se asocia con el pH?

```
lmer(largo ~ pH + (1 | poblacion/familia))
lme(largo ~ pH , random = ~ 1|poblacion/familia)
```

- ¿Y si se sospecha que el "efecto" del pH sobre el largo del ejemplar cambia entre poblaciones?

http://mfviz.com/hierarchical-models/

Modelos lineales mixtos



a) Modelo sin efectos aleatorios

b) Modelo con intercepto aleatorio

c) Modelo con pendiente aleatoria

d) Modelo con intercepto y pendiente aleatoria

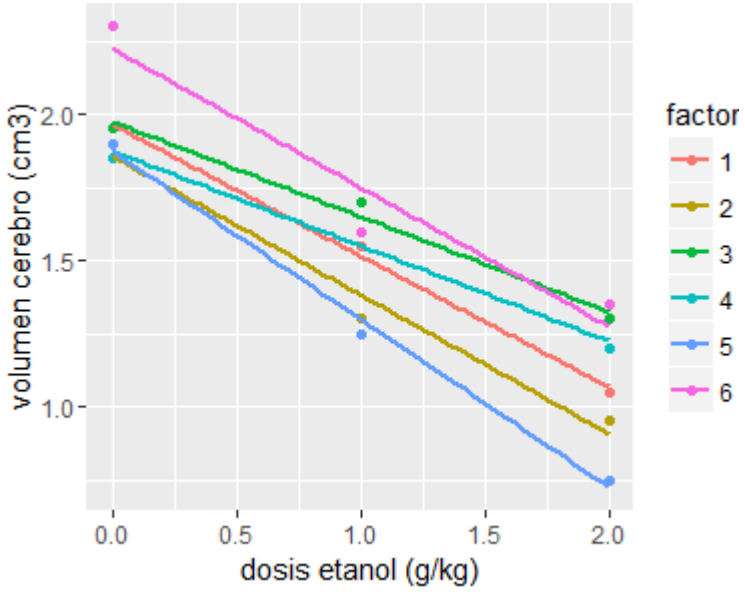
```
a <-lm(Y ~ X, data)
b <-lmer (Y ~ X + (1+|Factor_aleatorio), data)
c <-lmer (Y ~ X + (0+X|Factor_aleatorio), data)
d <-lmer (Y ~ X + (1+X|Factor_aleatorio),data)
```

Modelos con intercepto y pendiente aleatorios

- ¿Y si se sospecha que el "efecto" del pH sobre el largo del ejemplar entre poblaciones?

```
lmer(largo ~ pH + (1 + pH | poblacion))
lme(largo ~ pH , random = ~ 1 + pH |poblacion)
```

Implica una interacción trans-nivel



En el ej de DBA:

Implica interacción tratamiento x bloque

```
lmer(vol ~ etanol + (etanol| camada), bd)
```

Modelos con VE de efectos aleatorios cruzados

En un ensayo de comparabilidad interlaboratorios, se suministraron 5 muestras de suero de pacientes a 10 laboratorios. Cada laboratorio midió la concentración de un anticuerpo por triplicado

Se desea estudiar la variabilidad intra e interlaboratorios

```
lmer(Y ~ X + (1 | A) + (1 | B), data)
```

La clave está en el armado de la base de datos

Laboratorio	Paciente	Concentr
L1	1	12
L1	1	19
L1	1	23
L1	5	
L10	5	
L10	5	

Bibliografía

- Pinheiro J.C., Bates D.M. 2004. Mixed-Effects Models in S and S-PLUS. Springer, New York
- Zuur, A., Ieno, E.N., Walker, N., Saveliev, A.A., Smith, G.M. 2009. Mixed Effects Models and Extensions in Ecology with R. Springer, New York
- Zuur AF, Hilbe JM and Ieno EN. 2013. Beginner's Guide to GLM and GLMM with R . Highland Statistics Ltd
- Clark JS. 2006. Hierarchical modelling for the environmental sciences. Oxford University Press