```
> salida.gusanos<- glm(SF~sex*ldose, family=binomial)
> salida.sex<- glm(SF~sex, family=binomial)
> anova(salida.sex,salida.gusanos)

Analysis of Deviance Table

Model 1: SF ~ sex
Model 2: SF ~ sex * ldose
  Resid. Df Resid. Dev Df Deviance
1        10    118.799
2         8      4.994  2   113.81
```

Intermezzo: Bootstrap

Aproximación de la distribución de un estimador

- Laboratorio de Física I con 30 mesas.
- Cada grupo realiza el mismo experimento, obtiene n datos y calcula la estimación utilizando sus datos: **todos mismo** n
- Cada grupo obtiene una realización del estimador:

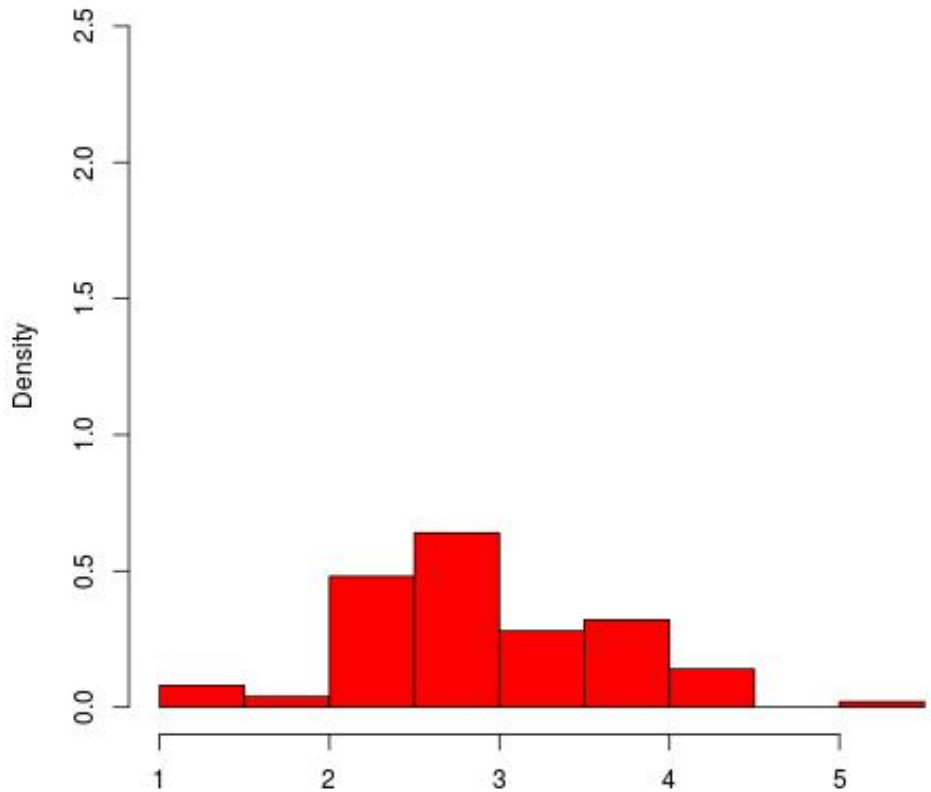
$$\hat{\theta}_1^n, \ldots, \hat{\theta}_{30}^n$$

- Cambio de notación. Desaparece n.

$$\hat{\theta}_1, \ldots, \hat{\theta}_{30}$$

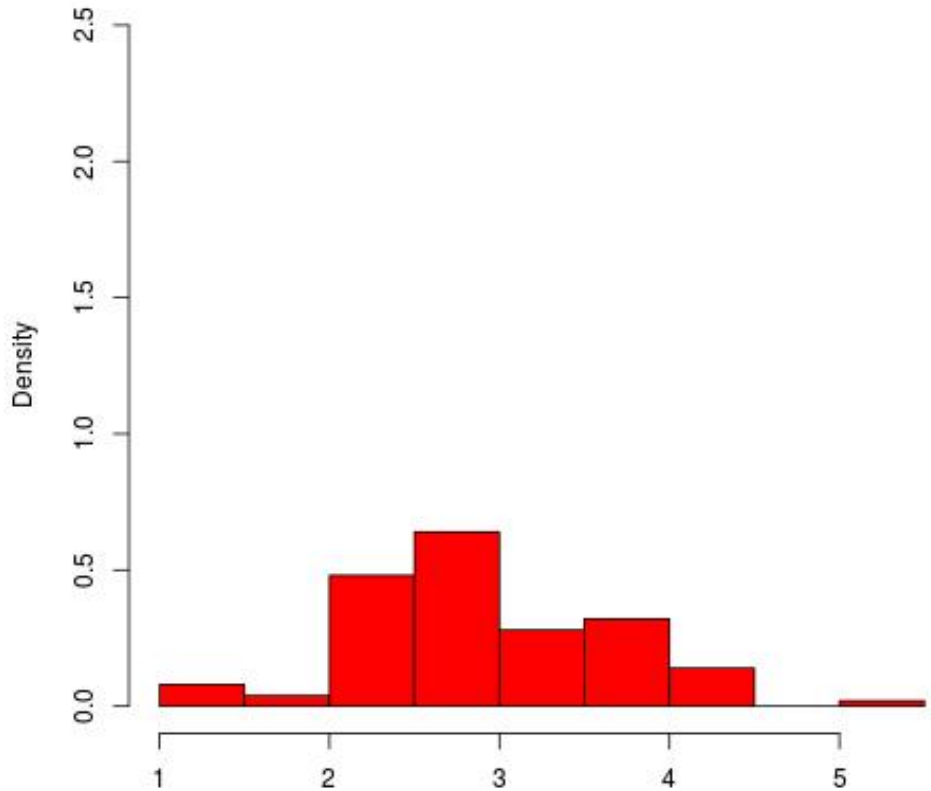
- Armamos un histograma con los valores obtenidos por cada grupo para aproximar la distribución del estimador.

Apoximación de la distribución de un estimador



Apoximación de la distribución de un estimador (si pudieramos...)

Situación muy especial!!!!: en general tenemos una sola realización del experimento



Apoximación de la distribución de un estimador

$$\hat{\theta}_1, \ldots, \hat{\theta}_{N_{Boot}} \ , \quad U_i \sim F$$

Aproximación Bootstrap de la distribución de un estimador

$$U_1 \ldots, U_n \quad U_i \sim F \dashrightarrow \hat{\theta}_1, \ldots, \hat{\theta}_{N_{Boot}}$$

$u_1 \ldots, u_n$: datos originales, realizaciones de $U_i \sim F \dashrightarrow \tilde{F}_n$:

$\tilde{F}_n$: una distribución construida con los datos originales que aproxima a F

$$\hat{\theta}_1^*, \ldots, \hat{\theta}_{N_{Boot}}^* , \quad U_i^* \sim \tilde{F}_n$$

Por ejemplo, podemos tomar $\tilde{F}_n = \hat{F}_n$, siendo $\hat{F}_n$ la distribución empírica.

Estimadores Asintóticamente Normales

- $\widehat{\theta}_n$ se dice asintóticamente normal (A.N) sii

$$\frac{\widehat{\theta}_n - \theta}{\text{se}} \approx \mathcal{N}(0,1),$$

donde $\text{se} = \text{se}(\widehat{\theta}_n)$ denota el desvío estandar del estimador $\widehat{\theta}_n$.

- Ejemplo: $\widehat{\mu}_n = \bar{U}_n$ es a.n., por TCL.

$$\frac{\bar{U}_n - \mu}{\text{se}} \approx \mathcal{N}(0,1),$$

siendo

$$\text{se} = \text{se}(\bar{U}_n) = \sqrt{\mathbb{V}(\bar{U}_n)} = \sqrt{\frac{\sigma^2}{n}}$$

Intervalos de Confianza Asintóticamente Normal

- Sea $\widehat{\theta}_n$ asintóticamente normal

$$\frac{\widehat{\theta}_n - \theta}{\text{se}} = \frac{\widehat{\theta}_n - \theta}{\text{se}(\widehat{\theta}_n)} \approx \mathcal{N}(0,1)$$

- Sea $\widehat{\text{se}}$ tal que $\frac{\text{se}(\widehat{\theta}_n)}{\widehat{\text{se}}} \to 1$,
- Tenemos entonces que

$$\left(\widehat{\theta}_n - z_{\alpha/2}\,\widehat{\text{se}} \quad , \quad \widehat{\theta}_n + z_{\alpha/2}\,\widehat{\text{se}}\right)$$

es un intervalo de confianza de nivel asintótico $1-\alpha$ para θ.

Intervalo de confianza para la media

- $\mu := \mathbb{E}_F(U)$. Estimador: $\widehat{\mu}_n = \bar{U}_n$
- Distribución de $\widehat{\mu}_n$: asintóticamente normal

$$\frac{\widehat{\mu}_n - \mu}{\text{se}(\widehat{\mu}_n)} \approx \mathcal{N}(0,1) \quad n \text{ grande}$$

- Desvío del Estimador:

$$\text{se}(\widehat{\mu}) = \sqrt{\mathbb{V}_F(\widehat{\mu})} = \sqrt{\frac{\sigma^2}{n}} = \text{se}$$

se estima con $\widehat{\text{se}} = \sqrt{\frac{\widehat{\sigma}^2}{n}}$ o con $\widehat{\text{se}} = \sqrt{\frac{S^2}{n}}$

$$\text{Intervalo de confianza} \quad \widehat{\mu} \pm z_{\alpha/2}\widehat{\text{se}}$$

¿Intervalo de confianza para la mediana?

- Distribución de la mediana muestral: asintóticamente normal

$$\frac{\text{med}(U_1, \ldots, U_n) - \text{med}(U)}{\text{se}} \approx \mathcal{N}(0,1) \quad n \text{ grande}$$

- Desvío del Estimador:

$$\text{se} = \text{se}(\text{med}(U_1, \ldots, U_n)) = \sqrt{\mathbb{V}_F\{\text{med}(U_1, \ldots, U_n)\}} = ???$$

- $\widehat{\text{se}} = ??$
- Bootstrap! $\widehat{\text{se}}_{boot}$

$$\text{Intervalo de confianza} \quad \text{med}(U_1, \ldots, U_n) \pm z_{\alpha/2}\widehat{\text{se}}_{boot}$$

Intevalos Bootstrap Normal

- $\widehat{\theta}_n$ asintóticamente normal si

$$\frac{\widehat{\theta}_n - \theta}{\text{se}} \approx \mathcal{N}(0,1)$$

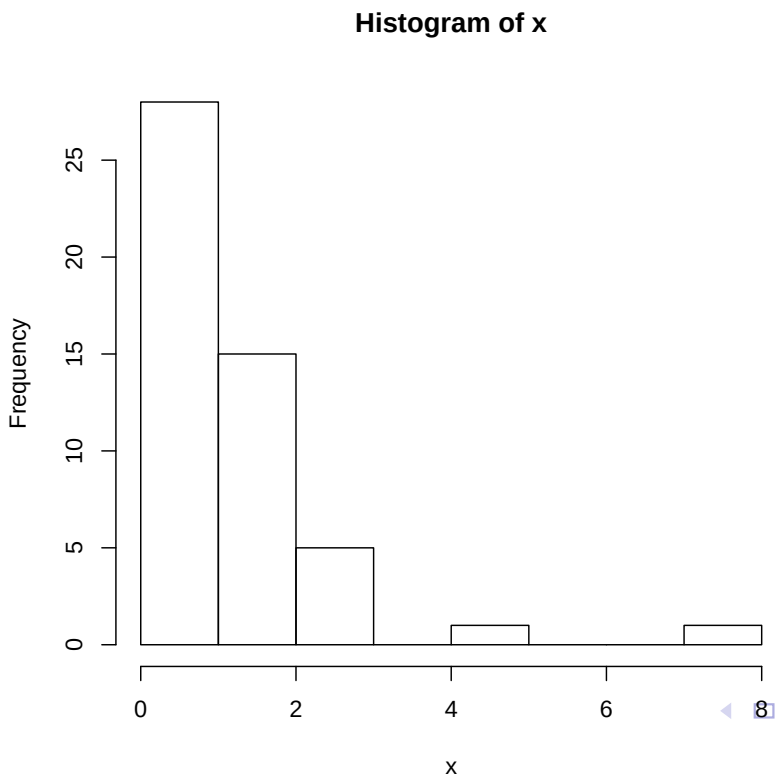
  con $\text{se} = \text{se}(\widehat{\theta}_n)$
- Sea $\widehat{\text{se}}_{\text{boot}}$ el estimador bootstrap de $\text{se}(\widehat{\theta}_n)$

  intervalo boot normal nivel $1-\alpha$: $\quad \widehat{\theta}_n \pm z_{\alpha/2}\, \widehat{\text{se}}_{\text{boot}}$

Toy Example

```
> set.seed(123)
> ene<- 50
> x<- rexp(ene)
> mediana<-median(x)
> mediana

[1] 0.8445896
```



Bootstrap: ejemplo mediana

```
>   boot=2000
>   median.boot<- rep(NA,boot)
> #
>   set.seed(999)
> #
> #
>   for(i in 1:boot)
+   {
+     ind.boot<-sample(1:ene, replace = TRUE)
+     median.boot[i]<-median(x[ind.boot])
+   }
```

Bootstrap: ejemplo mediana

```
>   head(median.boot)

[1] 0.6135649 0.8497861 0.9057966 0.9853506 0.8434573 0.7906818

>   median.prom.boot<- mean(median.boot)
>   var.median.boot<- mean((median.boot-median.prom.boot)^2)
> #
>   alfa= 0.05
>   limites.median_LI<- mediana-qnorm(1-alfa/2)*sqrt(var.median.boot)
>   limites.median_LS<- mediana+qnorm(1-alfa/2)*sqrt(var.median.boot)
>   c(limites.median_LI,limites.median_LS)

[1] 0.5359824 1.1531968

>
```

Intevalos Bootstrap Percentil

- N_{boot}
- Sean $\widehat{\theta}^*_1, \ldots, \widehat{\theta}^*_{Nboot}$ estadísticos bootstrap de su estimador.

$$\text{intervalo boot percentil } 1-\alpha\text{:} \quad \left(\widehat{\theta}^*_{\alpha/2}\ ,\ \widehat{\theta}^*_{1-\alpha/2}\right)$$

Bootstrap: ejemplo mediana

```
>   head(median.boot)

[1] 0.6135649 0.8497861 0.9057966 0.9853506 0.8434573 0.7906818

>   #
>   alfa= 0.05
> ##
>   IC.median.per_LI<-  quantile(median.boot, probs = alfa/2, na.rm = T)
>   IC.median.per_LS<-  quantile(median.boot, probs = 1-alfa/2, na.rm = T)
> ##
>   c(IC.median.per_LI,IC.median.per_LS)

     2.5%     97.5%
0.5826475 1.1775316

>
```

Bootstrap en Regresión Logística (RL)

En regresión y en particular en RL, hay que tener algunos cuidados al bootstrapear.

El primero es que el EMV para RL podría no existir para una muestra dada, ya que la condición para que en este modelo el EMV exista, es que debe haber *superposición* (overlapping) (Silvapulle, 1981 y Albert y Anderson, 1984).

Hay distintas maneras de realizar bootstrap en RL, en forma paramétrica y no paramétrica. Dos de ellas son las siguientes:

Boot NP : bootstrapear los pares $(x_1, y_1)^*, \ldots, (x_n, y_n)^*$, evaluar si existe el EMV, de lo contarario descartar y generar otra.

Boot P : Con la muestra original $(x_1, y_1), \ldots, (x_n, y_n)$, estimar $\widehat{\beta}$, generar muestras $(x_1^*, y_1^*), \ldots, (x_n^*, y_n^*)$, donde $x_i^* = x_i$ y cada respuesta y_1^* es generada siguiendo el modelo $p(x_i^*, \widehat{\beta})$. Evaluar si existe el EMV, de lo contarario, descartar y generar una nueva.

Otro Toy Example

```
> attach(toyexample)
> plot(dose,y)
> salida<-glm(y~dose, family=binomial())
> pp<- summary(salida)
> beta_0<- pp$coef[1,1]
> beta_1<- pp$coef[2,1]
> alfa<- 0.05
> #Intervalo Asintotico beta_0 nivel=1-alpha=0.95
> IC.AS_beta_0_LI<-beta_0-qnorm(1-alfa/2)*pp$coef[1,2]
> IC.AS_beta_0_LS<-beta_0+qnorm(1-alfa/2)*pp$coef[1,2]
> IC.AS_beta_0<- c(IC.AS_beta_0_LI,IC.AS_beta_0_LS)
> IC.AS_beta_0

[1] -6.282708 -2.357711

> #Intervalo Asintoticobeta_1 nivel=1-alpha=0.95
> IC.AS_beta_1_LI<-beta_1-qnorm(1-alfa/2)*pp$coef[2,2]
> IC.AS_beta_1_LS<-beta_1+qnorm(1-alfa/2)*pp$coef[2,2]
> IC.AS_beta_1<- c(IC.AS_beta_1_LI,IC.AS_beta_1_LS)
> IC.AS_beta_1

[1] 1.500288 3.585808
```

Opción NP

```
>   boot=2000
>   beta_0.boot<- beta_1.boot<- rep(NA,boot)
>   ncon<-0
>   n=length(y) # tamaño muestral
>   set.seed(999)
>   i<-1
>   while(i <= boot)
+   {
+     ind.boot<-sample(1:n, replace = TRUE)
+     saida<-glm(y[ind.boot]~dose[ind.boot],family=binomial)
+     if(saida$converged==FALSE)
+     {  ncon<-ncon+1
+        cat("No convergence at least of one B replicate\n")
+        next
+     }
+     beta_0.boot[i]<- saida$coef[1]
+     beta_1.boot[i]<- saida$coef[2]
+     i<- i+1
+   }
```

Opción NP

```
>   mean.beta_0<- mean(beta_0.boot)
>   var.beta_0.boot<- mean((beta_0.boot- mean.beta_0)^2)
>   limites.beta_0_LI<- beta_0-qnorm(1-alfa/2)*sqrt(var.beta_0.boot)
>   limites.beta_0_LS<- beta_0+qnorm(1-alfa/2)*sqrt(var.beta_0.boot)
>   IC.BN.beta_0<- c(limites.beta_0_LI,limites.beta_0_LS)
> ##
>   mean.beta_1<- mean(beta_1.boot)
>   var.beta_1.boot<- mean((beta_1.boot- mean.beta_1)^2)
>   limites.beta_1_LI<- beta_1-qnorm(1-alfa/2)*sqrt(var.beta_1.boot)
>   limites.beta_1_LS<- beta_1+qnorm(1-alfa/2)*sqrt(var.beta_1.boot)
>   IC.BN.beta_1<- c(limites.beta_1_LI,limites.beta_1_LS)
> ##
>   IC.Bper.beta_0<- unname(quantile(beta_0.boot, probs = c(alfa/2, 1-(alfa/2)),
+                                    na.rm = T))
> ##
>   IC.Bper.beta_1<- unname(quantile(beta_1.boot, probs = c(alfa/2, 1-(alfa/2)),
+                                    na.rm = T))
```

Opción NP

```
> ##
>   IC.AS_beta_0

[1] -6.282708 -2.357711

>   IC.BN.beta_0

[1] -6.440258 -2.200161

>   IC.Bper.beta_0

[1] -6.883286 -2.866992

> ##
>   IC.AS_beta_1

[1] 1.500288 3.585808

>   IC.BN.beta_1

[1] 1.373009 3.713087

>   IC.Bper.beta_1

[1] 1.765358 4.063044
```

Opción P

```
>     logistica<- function(u,beta0,beta1){
+       uu<- 1/(1+exp(-beta0-beta1*u))
+       return(uu)
+   }
> ##
> ##
>   probas.est<- rep(NA,n)
>   for (i in 1:n) {
+     probas.est[i]<- logistica(dose[i],beta_0,beta_1)
+   }
```

Opción P

```
>   boot=2000
>   beta_0.boot<- beta_1.boot<- rep(NA,boot)
>   ncon<-0
>   n=length(y) # tamaño muestral
>   x.star<- dose
>   set.seed(999)
>   i<-1
>   while(i <= boot)
+   {
+     y.star<-rbinom(n,1,probas.est)
+     saida<-glm(y.star~x.star,family=binomial)
+     if(saida$converged==FALSE)
+     {  ncon<-ncon+1
+     cat("No convergence at least of one B replicate\n")
+     next
+     }
+     beta_0.boot[i]<- saida$coef[1]
+     beta_1.boot[i]<- saida$coef[2]
+     i<- i+1
+   }
```

Opción P

```
>   mean.beta_0<- mean(beta_0.boot)
>   var.beta_0.boot<- mean((beta_0.boot- mean.beta_0)^2)
>   #
>   limites.beta_0_LI<- beta_0-qnorm(1-alfa/2)*sqrt(var.beta_0.boot)
>   limites.beta_0_LS<- beta_0+qnorm(1-alfa/2)*sqrt(var.beta_0.boot)
>   IC.BN.beta_0<- c(limites.beta_0_LI,limites.beta_0_LS)
> ##
> ##
>   mean.beta_1<- mean(beta_1.boot)
>   var.beta_1.boot<- mean((beta_1.boot- mean.beta_1)^2)
>   #
>   limites.beta_1_LI<- beta_1-qnorm(1-alfa/2)*sqrt(var.beta_1.boot)
>   limites.beta_1_LS<- beta_1+qnorm(1-alfa/2)*sqrt(var.beta_1.boot)
>   IC.BN.beta_1<- c(limites.beta_1_LI,limites.beta_1_LS)
>   IC.Bper.beta_0<- unname(quantile(beta_0.boot, probs = c(alfa/2, 1-(alfa/2)),
+                                    na.rm = T))
> ##
>   IC.Bper.beta_1<- unname(quantile(beta_1.boot, probs = c(alfa/2, 1-(alfa/2)),
+                                    na.rm = T))
```

Opción NP

```
> ##
>   IC.AS_beta_0

[1] -6.282708 -2.357711

>   IC.BN.beta_0

[1] -6.504139 -2.136279

>   IC.Bper.beta_0

[1] -7.130393 -2.765905

> ##
>   IC.AS_beta_1

[1] 1.500288 3.585808

>   IC.BN.beta_1

[1] 1.366014 3.720082

>   IC.Bper.beta_1

[1] 1.761656 4.066195
```