

# Tarea ensayos clínicos

*Gaëlle Cordier*

Se trata de analizar los datos de dos ensayos sencillos en los que se comparan dos grupos, en el primer caso mediante una respuesta continua y en el segundo mediante una respuesta dicotómica.

En ambos casos hay que comparar las respuestas de los dos grupos primero sin ajustes de ningún tipo y luego ajustando por las covariables que se consideren oportunas. También hay que seleccionar un tamaño muestral para un ensayo posterior algo mayor y redactar un pequeño informe con las conclusiones (que contenga un anexo con el script de R que se haya utilizado).

## PRIMER ENSAYO

Son datos de un ensayo clínico con pacientes con adenomas en colon y recto. (Giardiello et al., 1993, Treatment of colonic and rectal adenomas with Sulindac in familial adenomatous polyposis. New England Journal of Medicine, 328, 1313-1316).

Los datos muestran el número de pólipos, en logaritmos decimales, al principio del ensayo y al año de tratamiento:

grupo	antes	despues
tratado	0.84510	0.60206
control	0.69897	1.41497
tratado	1.36173	1.20412
control	1.54407	1.60206
tratado	1.04139	1.14613
control	1.07918	1.20412
control	0.84510	1.04139
control	2.50243	2.63749
tratado	2.20412	1.41497
tratado	0.90309	0.84510
control	1.30103	1.65321
control	1.04139	1.50515
control	1.38021	1.90309
tratado	1.53148	1.53148
control	1.73239	1.57978
control	1.47712	1.75587
tratado	1.00000	0.84510
tratado	1.30103	0.00000
tratado	1.07918	0.90309

Table 1: Datos del ensayo

Sumario de los datos:

grupo	variable	value
control:20	antes :19	Min. :0.000
tratado:18	despues:19	1st Qu.:1.010
		Median :1.301
		Mean :1.307
		3rd Qu.:1.541
		Max. :2.637

Table 2: Sumario

grupo	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
antes	0.699	1.021	1.301	1.309	1.504	2.502
despues	0.000	0.972	1.415	1.305	1.591	2.637

Table 3: Sumario antes/después

grupo	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
control	0.699	1.173	1.491	1.495	1.673	2.637
tratado	0.000	0.860	1.060	1.098	1.347	2.204

Table 4: Sumario por grupo de tratamiento

grupo	variable	n	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
control	antes	10	0.699	1.051	1.341	1.360	1.527	2.502
control	despues	10	1.041	1.438	1.591	1.630	1.730	2.637
tratado	antes	9	0.845	1.000	1.079	1.252	1.362	2.204
tratado	despues	9	0.000	0.845	0.903	0.944	1.204	1.531

Table 5: Sumario por grupo de tratamiento antes/después

## Objetivos del análisis estadístico:

### 1. Análisis estadístico de los datos. ¿Tiene algún efecto el tratamiento en los resultados?

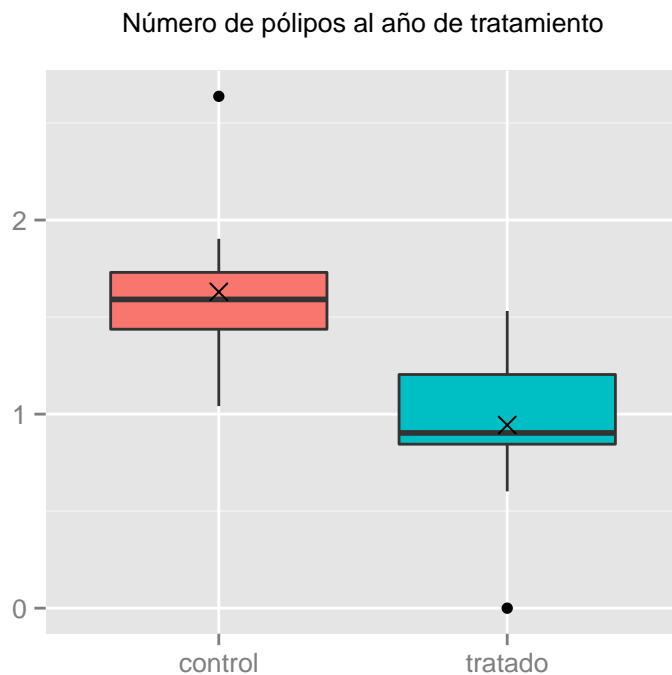
#### a) Primera aproximación: análisis de única variable respuesta despues

- Análisis descriptivo

grupo	n	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	S	SE
control	10	1.041	1.438	1.591	1.630	1.730	2.637	0.434	0.137
tratado	9	0.000	0.845	0.903	0.944	1.204	1.531	0.462	0.154

Table 6: Sumario por grupo de tratamiento antes/después

donde desviación típica  $S = \frac{s - \bar{x}}{n-1}$  y error estándar de la media  $SE_{\bar{x}} = \frac{s}{\sqrt{n}}$



El número de pólipos al año de tratamiento parece ser mayor en el grupo control que en el grupo tratado.

- **Análisis estadístico:** ¿es la media en el número de pólipos al año de tratamiento significativamente diferente entre los grupos control y tratado?

$$H_0 : \mu_c = \mu_t ; H_1 : \mu_c \neq \mu_t$$

$$\text{ANOVA: F-test} \rightarrow F = \frac{MS_B}{MSE}$$

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
grupo	1	2.230	2.23	11.138	0.004
Residuals	17	3.404	0.20	NA	NA

Table 7: ANOVA control-tratamiento

donde:

- **Sum Sq** es la suma de cuadrados entre grupos  $SSB$  (**grupo**) y la suma de cuadrados dentro de los grupos  $SSE$  (**residuals**)
- **Mean Sq** es la variabilidad entre grupos  $MS_B$  (**grupo**) y la variabilidad dentro de los grupos  $MSE$  (**residuals**)
- **F value** es el F-ratio (**grupo**)
- **Pr(>F)** es el p-value (**grupo**)

F-ratio  $> 1$  es significativo  $\rightarrow MS_B > MSE \rightarrow$  se rechaza  $H_0$

$$\text{Comparación de medias: t-test} \rightarrow t = \frac{\bar{x}_t - \bar{x}_c}{SE_{\text{diff}}}$$

$$\text{donde } SE_{\text{diff}} = \sqrt{\frac{S_c^2}{n_c} + \frac{S_t^2}{n_t}} = \frac{S_c}{\sqrt{n_c}} + \frac{S_t}{\sqrt{n_t}}$$

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	1.630	0.141	11.518	0.000
grupotratado	-0.686	0.206	-3.337	0.004

Table 8: t-test control-tratamiento

donde:

- **Estimate** es la estimación de la media  $\bar{x}_c$  (**Intercept**) y la estimación de la diferencia entre las medias  $\bar{x}_t - \bar{x}_c$  (**grupotratado**)
- **Std. Error** es el error estándar de la media  $SE_{\bar{x}_c}$  (**Intercept**) y el error estándar de la diferencia entre las medias  $SE_{\text{diff}}$  (**grupotratado**)
- **|t-value| = t** (**grupotratado**)
- **Pr(>|t|)** es el p-value (**grupotratado**)

$Pr(> |t|)$  es significativo  $\rightarrow \bar{x}_c - \bar{x}_t \neq 0 \rightarrow$  se rechaza  $H_0$

b) Segunda aproximación: análisis de variable efecto <- antes-despues

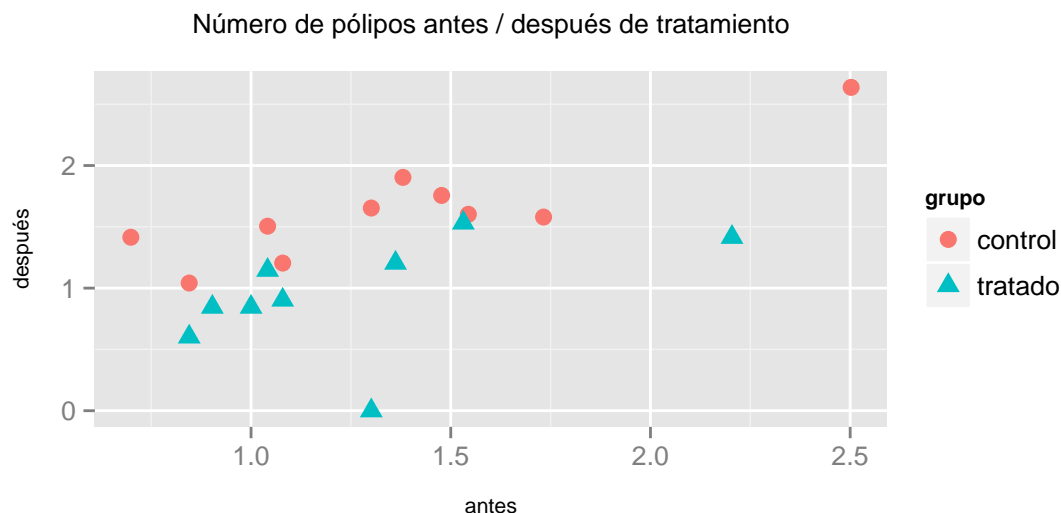
grupo	antes	despues	efecto
tratado	0.84510	0.60206	0.24304
control	0.69897	1.41497	-0.71600
tratado	1.36173	1.20412	0.15761
control	1.54407	1.60206	-0.05799
tratado	1.04139	1.14613	-0.10474
control	1.07918	1.20412	-0.12494
control	0.84510	1.04139	-0.19629
control	2.50243	2.63749	-0.13506
tratado	2.20412	1.41497	0.78915
tratado	0.90309	0.84510	0.05799
control	1.30103	1.65321	-0.35218
control	1.04139	1.50515	-0.46376
control	1.38021	1.90309	-0.52288
tratado	1.53148	1.53148	0.00000
control	1.73239	1.57978	0.15261
control	1.47712	1.75587	-0.27875
tratado	1.00000	0.84510	0.15490
tratado	1.30103	0.00000	1.30103
tratado	1.07918	0.90309	0.17609

Table 9: Datos del ensayo 2

- Análisis descriptivo

grupo	n	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	S	SE
control	10	-0.716	-0.436	-0.238	-0.270	-0.128	0.153	0.253	0.08
tratado	9	-0.105	0.058	0.158	0.308	0.243	1.301	0.449	0.15

Table 10: Sumario del efecto por grupo de tratamiento



Para mismos valores de **antes**, el grupo control tiene valores mayores de **después** → parece haber diferencias en el efecto entre los dos grupos de tratamiento.

- **Análisis estadístico:** ¿es la media del efecto significativamente diferente entre los grupos control y tratado?

$$H_0 : \mu_c = \mu_t ; H_1 : \mu_c \neq \mu_t$$

**ANOVA:** F-test  $\rightarrow F = \frac{MS_B}{MSE}$

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
grupo	1	1.582	1.582	12.287	0.003
Residuals	17	2.188	0.129	NA	NA

Table 11: ANOVA control-tratamiento

F-ratio  $> 1$  es significativo  $\rightarrow MS_B > MSE \rightarrow$  se rechaza  $H_0$

**Comparación de medias:** t-test  $\rightarrow t = \frac{\bar{x}_t - \bar{x}_c}{SE_{diff}}$

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-0.270	0.113	-2.375	0.030
grupotratado	0.578	0.165	3.505	0.003

Table 12: t-test control-tratamiento

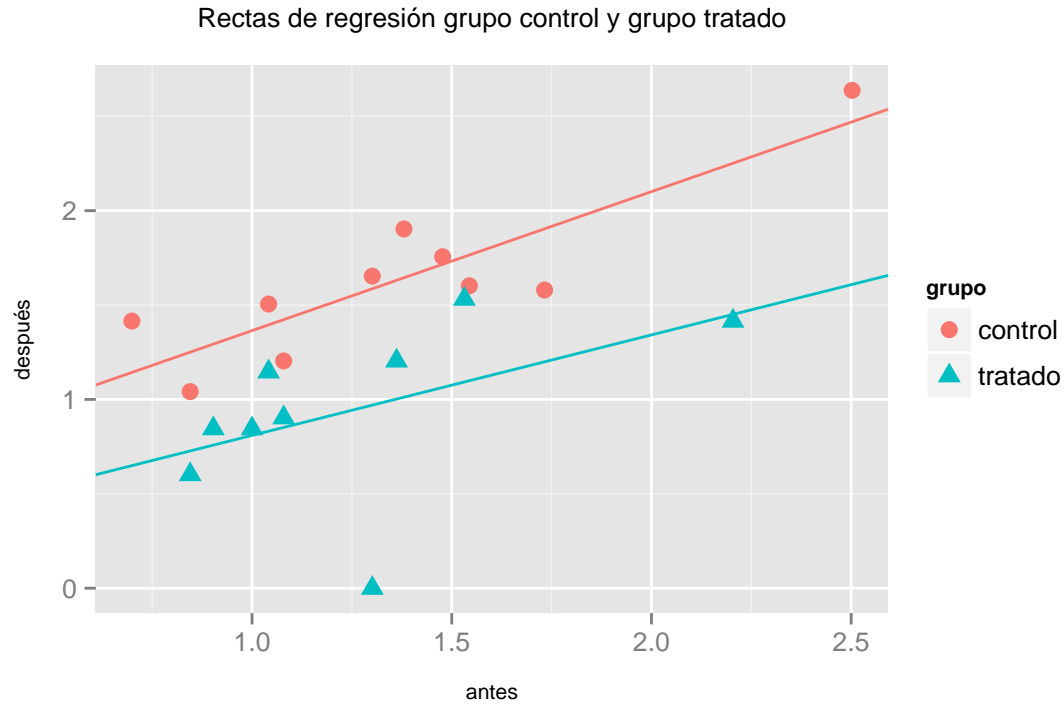
$Pr(> |t|)$  es significativo  $\rightarrow \bar{x}_c - \bar{x}_t \neq 0 \rightarrow$  se rechaza  $H_0$

### c) Tercera aproximación: ANCOVA (covariable antes)

- **Análisis descriptivo**

$$\text{despues}_c = \alpha_c + \beta_c * \text{antes}_c$$

$$\text{despues}_t = \alpha_t + \beta_t * \text{antes}_t$$



El intercept  $\alpha$  es diferente para cada grupo  $\rightarrow$  el grupo de tratamiento parece tener efecto sobre la variable respuesta **después**, donde el grupo control presentaría mayor número de pólipos al año de tratamiento que el grupo tratado.

La pendiente  $\beta$  es similar entre ambos grupos  $\rightarrow$  el efecto de la covariable **antes** sobre la variable respuesta **después** parece ser similar para ambos grupos, y no parece haber interacción entre la covariable y el grupo.

- **Análisis estadístico**

#### Ajuste de modelo completo con interacción:

$$\text{despues} = \alpha_{\text{grupo}} + \beta * \text{antes} + \alpha_{\text{grupo}}\beta * \text{antes}$$

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	0.628	0.317	1.985	0.066
grupotratado	-0.351	0.488	-0.719	0.483
antes	0.736	0.219	3.361	0.004
grupotratado:antes	-0.204	0.358	-0.570	0.577

Table 13: ANCOVA interacción

donde:

- **Intercept:** Intercept 1  $\alpha_{grupo=control} + \beta * (\text{antes}=x) + \alpha_{grupo=control}\beta * (\text{antes}=x)$
- **grupotratado:** diferencia Intercept 2  $\alpha_{grupo=tratado} - \text{Intercept 1}$
- **antes:** pendiente 1  $\beta * (\text{antes}=x+1)$
- **grupotratado:antes:** diferencia pendiente 2  $\beta * [(\text{grupo=tratado}):(\text{antes}=x+1)] - \text{pendiente 1}$

En presencia del término de interacción, los ‘efectos principales’ son efectos condicionales (hay tantos efectos **grupotratado** como niveles **antes**, y tantos efectos **antes** como niveles **grupotratado**).

El efecto debido al término de interacción **grupotratado:antes** no es significativo  $\rightarrow$  debe eliminarse del modelo.

Suma de cuadrados de tipo I (secuencial):

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
grupo	1	2.230	2.230	19.546	0.000
antes	1	1.655	1.655	14.507	0.002
grupo:antes	1	0.037	0.037	0.325	0.577
Residuals	15	1.711	0.114	NA	NA

Table 14: Tabla ANOVA interacción (A,B,AB)

donde **grupo:**  $SS(B)$ , **antes:**  $SS(A|B)$ , **grupo:antes:**  $SS(A * B|A, B)$

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
antes	1	2.121	2.121	18.592	0.001
grupo	1	1.764	1.764	15.461	0.001
antes:grupo	1	0.037	0.037	0.325	0.577
Residuals	15	1.711	0.114	NA	NA

Table 15: Tabla ANOVA interacción (B,A,AB)

donde: **antes:**  $SS(A)$ , **grupo:**  $SS(B|A)$ , **antes:grupo:**  $SS(A * B|A, B)$

La variabilidad debida a la interacción **grupo:antes** no es significativa  $\rightarrow$  análisis de efectos principales.



### Ajuste de modelo sin interacción:

despues =  $\alpha_{grupo} + \beta * \text{antes}$

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	0.732	0.253	2.892	0.011
grupotratado	-0.615	0.153	-4.018	0.001
antes	0.660	0.170	3.892	0.001

Table 16: ANCOVA efectos principales

donde:

- **Intercept:** Intercept 1  $\alpha_{grupo=control} + \beta * (\text{antes}=x)$
- **grupotratado:** diferencia Intercept 2  $\alpha_{grupo=tratado} - \text{Intercept 1}$
- **antes:** pendiente  $\beta * (\text{antes}=x+1)$

El factor **grupo** y la covariable **antes** son significativos.

Suma de cuadrados de tipo II:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
grupo	1	2.230	2.230	20.407	0.000
antes	1	1.655	1.655	15.146	0.001
Residuals	16	1.749	0.109	NA	NA

Table 17: Tabla ANOVA efectos principales (A,B)

donde **antes:**  $SS(B|A)$

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
antes	1	2.121	2.121	19.411	0.000
grupo	1	1.764	1.764	16.142	0.001
Residuals	16	1.749	0.109	NA	NA

Table 18: Tabla ANOVA efectos principales (B,A)

donde **grupo:**  $SS(A|B)$

	Df	Sum of Sq	RSS	AIC	F value	Pr(>F)
	NA	NA	1.748519	-39.32773	NA	NA
grupo	1	1.764084	3.512602	-28.07355	16.14243	0.0009942
antes	1	1.655198	3.403717	-28.67185	15.14606	0.0012960

Table 19: SSII efectos principales

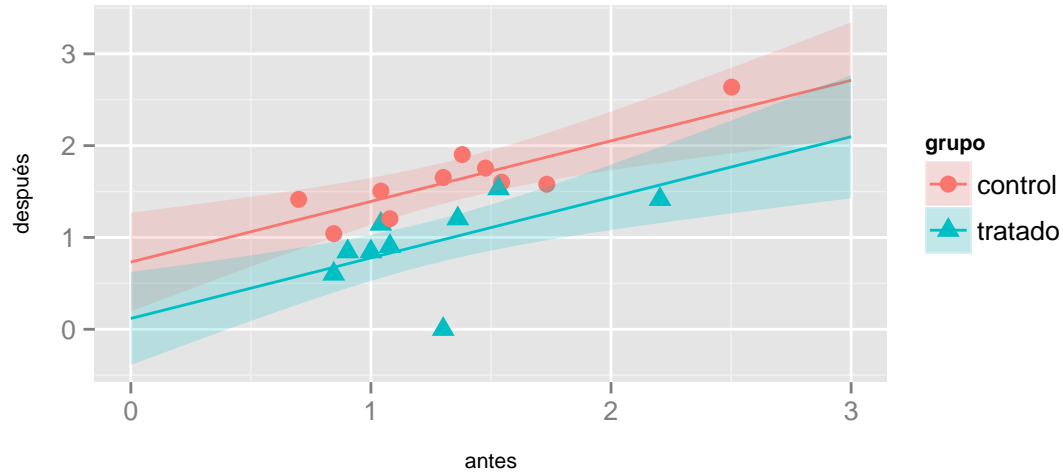
donde **grupo**:  $SS(A|B)$  y **antes**:  $SS(B|A)$

Ambos efectos principales **grupo** y **antes** son significativos en presencia del otro.

El modelo mínimo adecuado es:  $\text{despues} = \alpha_{\text{grupo}} + \beta * \text{antes}$

grupo	antes	despues	fit	IC_lwr	IC_upr	grupo	antes	despues	fit	IC_lwr	IC_upr
control	0.699	1.415	1.193	0.868	1.518	tratado	0.845	0.602	0.675	0.400	0.951
control	0.845	1.041	1.290	1.001	1.579	tratado	0.903	0.845	0.713	0.448	0.979
control	1.041	1.505	1.419	1.170	1.669	tratado	1.000	0.845	0.777	0.527	1.028
control	1.079	1.204	1.444	1.201	1.688	tratado	1.041	1.146	0.805	0.559	1.050
control	1.301	1.653	1.591	1.368	1.813	tratado	1.079	0.903	0.830	0.588	1.071
control	1.380	1.903	1.643	1.421	1.865	tratado	1.301	0.000	0.976	0.742	1.210
control	1.477	1.756	1.707	1.481	1.932	tratado	1.362	1.204	1.016	0.779	1.253
control	1.544	1.602	1.751	1.520	1.982	tratado	1.531	1.531	1.128	0.874	1.382
control	1.732	1.580	1.875	1.616	2.134	tratado	2.204	1.415	1.572	1.157	1.986
control	2.502	2.637	2.383	1.917	2.850	NA	NA	NA	NA	NA	NA

Table 20: Ajuste modelo final - IC 95%



2. Se va a realizar un ensayo algo más grande en el que se va a utilizar un nivel de significatividad  $\alpha = 0.05$ . Se desea obtener una potencia del 95% para detectar diferencias de medias (en escala logarítmica) de 0.4 unidades. ¿Cuántos pacientes por grupo se necesitan? ¿Variaría mucho el tamaño muestral si en lugar de utilizar un estimador puntual de la varianza, común a ambos grupos, se utilizara el extremo superior del intervalo de confianza al 80% sobre esa varianza desconocida?

- Utilizando estimador puntual de varianza común

$$\alpha = 0.05$$

$$\text{potencia} = 0.95$$

$$\delta = \mu_{t.\text{despues}} - \mu_{c.\text{despues}} = 0.4$$

$$n_c = 10$$

$$n_t = 9$$

$$S_c^2 = 0.19$$

$$S_t^2 = 0.21$$

$$S_{x_c x_t} = \sqrt{\frac{(n_c-1)s_{x_c}^2 + (n_t-1)s_{x_t}^2}{n_c + n_t - 2}} = 0.45$$

delta	sd	alpha	power	N
0.4	0.45	0.05	0.95	34

Table 21: Two-sample t test power calculation

Tamaño muestral necesario para cada grupo  $N = 34$

ó

$$\alpha = 0.05$$

$$\beta = 1 - \text{potencia} = 1 - 0.95 = 0.05$$

$$\delta = \mu_{t.\text{despues}} - \mu_{c.\text{despues}} = 0.4$$

$$S_{x_c x_t} = 0.45$$

$$N = \frac{2\sigma^2(Z_{1-\beta} + Z_{1-\alpha/2})^2}{(\mu_t - \mu_c)^2}$$

Tamaño muestral necesario para cada grupo  $N = 33$

- Utilizando extremo superior del intervalo de confianza al 80% sobre varianza desconocida

¿?

## **Informe final**

blablabla

## ANEXO I

### Datos

```
grupo <- as.factor(c(1,0,1,0,1,0,0,0,1,1,0,0,0,1,0,0,1,1,1))
levels(grupo) <- c('control','tratado')
dat.polipos <- data.frame(grupo= grupo,
                          antes = c(0.84510,0.69897,1.36173,1.54407,1.04139,1.07918, 0.84510,
                                     2.50243,2.20412,0.90309,1.30103,1.04139,1.38021,1.53148,
                                     1.73239,1.47712,1.00000, 1.30103,1.07918),
                          despues=c(0.60206,1.41497,1.20412,1.60206,1.14613,1.20412,1.04139,
                                    2.63749,1.41497,0.84510,1.65321,1.50515,1.90309,1.53148,
                                    1.57978,1.75587,0.84510,0.00000,0.90309))

# tabla datos
kable(dat.polipos, caption = "Datos del ensayo")

# datos long-format
dat.pol_lf<-melt(data = dat.polipos,measure.vars = c("antes","despues"))

# resumen
sum1<-summary(dat.pol_lf)
# para sustituir NAs por "" en grupo y variable en `kable`:
sum1<-data.frame(grupo=c(sum1[1:2,1],rep("",4)),
                  variable=c(sum1[1:2,2],rep("",4)),
                  value=sum1[,3])
kable(sum1,caption = "Resumen")

# resumen antes/después
sum_names<-c("grupo","Min.","1st Qu.","Median","Mean","3rd Qu.","Max.")
kable(summaryBy(formula = value~variable,data = dat.pol_lf,FUN = summary),
      caption = "Resumen antes/después",
      col.names = sum_names,
      digits = 3)

# resumen grupo
kable(summaryBy(formula = value~grupo,data = dat.pol_lf,FUN = summary),
      caption = "Resumen por grupo de tratamiento",
      col.names = sum_names,
      digits = 3)

# resumen grupo antes/después
sum2<-summaryBy(formula = value~grupo*variable,data = dat.pol_lf,FUN = summary)
# summaryBy no tiene `n`:
n<-summaryBy(formula = value~grupo*variable,data = dat.pol_lf,FUN = length)
sum2_n<-data.frame(sum2[,1:2],n$value.length,sum2[,3:8])
sum_names_n<-c("grupo","variable","n","Min.","1st Qu.","Median","Mean","3rd Qu.","Max.")
kable(sum2_n,
      caption = "Resumen por grupo de tratamiento antes/después",
      col.names = sum_names_n,
      digits = 3)
```

## a) Primera aproximación

- Análisis descriptivo

```
# sumario
sum_d<-summaryBy(formula = despues~grupo,data = dat.polipos,FUN = summary)
# + n:
n_d<-summaryBy(formula = despues~grupo,data = dat.polipos,FUN = length)[,2]
# + s y se
sd_despues<-summaryBy(formula = despues~grupo,data = dat.polipos,FUN = sd)[,2]
se_despues<-sd_despues/sqrt(n_d)
# tabla
sum_d_ns<-data.frame(sum_d[,1],n_d,sum_d[,2:7],sd_despues,se_despues)
sum_d_names<-c("grupo","n","Min.","1st Qu.","Median","Mean","3rd Qu.","Max.","S","SE")
kable(sum_d_ns,col.names = sum_d_names,
      caption = "Sumario por grupo de tratamiento antes/después",
      digits = 3)
```

```
# boxplot
ggplot(data = dat.polipos, aes(x = grupo, y=despues, fill=grupo))+
  geom_boxplot()+
  stat_summary(fun.y="mean", geom="point", shape=4, size=3)+
  labs(x="",y="",title="Número de pólipos al año de tratamiento\n")+
  theme(plot.title=element_text(size=10))+
  scale_fill_discrete(guide=F)
```

- Análisis estadístico

```
# test heterocedasticidad
bartlett.test(formula = despues~grupo,data = dat.polipos)
```

```
##
## Bartlett test of homogeneity of variances
##
## data: despues by grupo
## Bartlett's K-squared = 0.0294, df = 1, p-value = 0.8638
```

```
# --> hay homogeneidad de varianzas entre los grupos control y tratamiento
```

```
# ANOVA
despues.aov<-summary(aov(formula = despues~grupo,data = dat.polipos))
kable(despues.aov[[1]],caption = "ANOVA control-tratamiento",digits = 3)
```

```
# t-test
despues.lm<-summary(lm(formula = despues~grupo,data = dat.polipos))
kable(despues.lm$coefficients,
      caption = "t-test control-tratamiento",
      digits = 3)
```

## b) Segunda aproximación

```
# datos efecto
dat.polipos_effect<-data.frame(dat.polipos,efecto=dat.polipos$antes-dat.polipos$despues)
kable(dat.polipos_effect,caption = "Datos del ensayo 2")
```

- Análisis descriptivo

```
# resumen
sum_e<-summaryBy(formula = efecto~grupo,data = dat.polipos_effect,FUN = summary)
# + n
n_e<-summaryBy(formula = efecto~grupo,data = dat.polipos_effect,FUN = length)[,2]
# + sd y se
sd_e<-summaryBy(formula = efecto~grupo,data = dat.polipos_effect,FUN = sd)[,2]
se_e<-sd_e/sqrt(n_e)

# tabla
sum_e_ns<-data.frame(sum_e[,1],n_e,sum_e[,2:7],sd_e,se_e)
sum_e_names<-c("grupo","n","Min.","1st Qu.","Median","Mean","3rd Qu.","Max.","S","SE")
kable(sum_e_ns,col.names = sum_e_names,
      caption = "Sumario del efecto por grupo de tratamiento",
      digits = 3)
```

```
# plot
ggplot(data = dat.polipos,aes(x = antes,y = despues,colour = grupo,shape = grupo))+
  geom_point(size=3)+
  labs(title="Número de pólipos antes / después de tratamiento\n",
       x="\nantes",y="después\n")+
  theme(plot.title=element_text(size=10),
        axis.title=element_text(size=8),
        legend.title=element_text(size=8))
```

- Análisis estadístico

```
# test heterocedasticidad
bartlett.test(formula = efecto~grupo,data = dat.polipos_effect)
```

```
##
## Bartlett test of homogeneity of variances
##
## data:  efecto by grupo
## Bartlett's K-squared = 2.5576, df = 1, p-value = 0.1098
```

```
# --> hay homogeneidad de varianzas entre los grupos control y tratamiento
```

```
# ANOVA
efecto.aov<-summary(aov(formula = efecto~grupo,data = dat.polipos_effect))
kable(efecto.aov[[1]],caption = "ANOVA control-tratamiento",digits = 3)
```

```
# t-test
efecto.lm<-summary(lm(formula = efecto~grupo,data = dat.polipos_effect))
kable(efecto.lm$coefficients,
      caption = "t-test control-tratamiento",
      digits = 3)
```

### c) Tercera aproximación

- Análisis descriptivo

```
# ajuste grupo control
cont.lm<-with(dat.polipos,lm(despues[grupo=="control"]~antes[grupo=="control"]))
cont.coefs<-cont.lm$coefficients

# ajuste grupo tratado
trat.lm<-with(dat.polipos,lm(despues[grupo=="tratado"]~antes[grupo=="tratado"]))
trat.coefs<-trat.lm$coefficients

grupo.coefs<-data.frame(grupo=c("control","tratado"),
                        Intercept=c(cont.coefs[1],trat.coefs[1]),
                        slope=c(cont.coefs[2],trat.coefs[2]))

# plot
ggplot(data = dat.polipos,aes(x = antes,y = despues,colour = grupo,shape = grupo))+
  geom_point(size=3)+
  labs(title="Rectas de regresión grupo control y grupo tratado\n",
       x="\nantes",y="después\n")+
  theme(plot.title=element_text(size=10),
        axis.title=element_text(size=8),
        legend.title=element_text(size=8))+
  geom_abline(data=grupo.coefs,aes(intercept=Intercept,slope=slope,colour=grupo))
```

- Análisis estadístico

```
# ajuste
pol.lm_max<-with(dat.polipos,lm(despues~grupo*antes))
ancova<-summary(pol.lm_max)
# summary
kable(ancova$coefficients,
      caption = "ANCOVA interacción",
      digits = 3)
```

```
# SSI (A,B,AB)
kable(anova(pol.lm_max),
      caption = "Tabla ANOVA interacción (A,B,AB)",
      digits = 3)
```

```
# SSI (B,A,AB)
pol.lm_max2<-with(dat.polipos,lm(despues~antes*grupo))
kable(anova(pol.lm_max2),
      caption = "Tabla ANOVA interacción (B,A,AB)",
      digits = 3)
```



```

# SSII (A,B)
kable(anova(pol.lm_main),
      caption = "Tabla ANOVA efectos principales (A,B)",
      digits = 3)

# SSII (B,A)
pol.lm_main2<-with(dat.polipos,lm(despues~antes+grupo))
kable(anova(pol.lm_main2),
      caption = "Tabla ANOVA efectos principales (B,A)",
      digits = 3)

# SSII
kable(drop1(pol.lm_main, ~ ., test="F"),caption = "SSII efectos principales")

# modelo mínimo adecuado
step(pol.lm_max)

## Start:  AIC=-37.73
## despues ~ grupo * antes
##
##           Df Sum of Sq    RSS    AIC
## - grupo:antes 1   0.03708 1.7485 -39.328
## <none>                        1.7114 -37.735
##
## Step:  AIC=-39.33
## despues ~ grupo + antes
##
##           Df Sum of Sq    RSS    AIC
## <none>                        1.7485 -39.328
## - antes  1      1.6552 3.4037 -28.672
## - grupo  1      1.7641 3.5126 -28.074

##
## Call:
## lm(formula = despues ~ grupo + antes)
##
## Coefficients:
## (Intercept) grupotratado      antes
##          0.7322      -0.6147      0.6598

# fitted data con IC
IC.fit<-predict(pol.lm_main,interval = "confidence") # level = 0.95 por defecto
dat.pol.fit<-data.frame(dat.polipos,fit=pol.lm_main$fitted.values,
                        IC_lwr=IC.fit[,2],IC_upr=IC.fit[,3])

# IC tabla
dat.pol.cont.IC<-dat.pol.fit[dat.pol.fit$grupo=="control",]
dat.pol.cont.IC2<-dat.pol.cont.IC[order(dat.pol.cont.IC$fit),]
dat.pol.trat.IC<-rbind(dat.pol.fit[dat.pol.fit$grupo=="tratado",],rep(NA,6))
dat.pol.trat.IC2<-dat.pol.trat.IC[order(dat.pol.trat.IC$fit,na.last = T),]
dat.pol.IC<-cbind(dat.pol.cont.IC2,dat.pol.trat.IC2)
row.names(dat.pol.IC)<-NULL
kable(dat.pol.IC,caption="Ajuste modelo final - IC 95%",digits=3)

```

```
# plot
ggplot(data = dat.pol.fit,aes(x = antes,y = despues,colour = grupo,shape = grupo))+
  geom_point(size=3)+
  labs(title="",x="\nantes",y="después\n")+
  theme(axis.title=element_text(size=8),
        legend.title=element_text(size=8))+
  geom_line(aes(x=antes,y=fit))+
  geom_ribbon(aes(y=fit,ymin=IC_lwr,ymax=IC_upr,fill=grupo),alpha=0.2,linetype="dashed")
```

## Tamaño muestral

- Estimador varianza

```
# desviación típica común
nc<-10
nt<-9
(var_c<-with(dat.polipos,var(despues[grupo=="control"])))
(var_t<-with(dat.polipos,var(despues[grupo=="tratado"])))
(sigm<-sqrt(((nc-1)*var_c+(nt-1)*var_t)/(nc+nt-2)))
# parámetros
alf<-0.05
delt<-0.4
pot<-0.95
```

```
# tamaño muestral

N<-power.t.test(delta = delt,sd = sigm,sig.level = alf,power = pot)
# (alternative="two.sided" por defecto)

kable(data.frame(delta=N$delta,sd=round(N$sd,2),alpha=N$sig.level,
               power=N$power,N=round(N$n,0)),
      caption="Two-sample t test power calculation")
```

```
# tamaño muestral 2
(bet<-1-pot)
(N2 <- round(2*(sigm/delt)^2*(sum(qnorm(c(1-bet,1-alf/2),0,1)))^2,0))
```

## SEGUNDO ENSAYO

Vamos a utilizar los datos `linfoma.dat` ya estudiados en la tarea sobre supervivencia. Ahora vamos a suponer que el objetivo principal del ensayo era la supervivencia a un horizonte de 1, no las curvas de supervivencia.

La variable respuesta, por tanto, será si el tiempo `B3TODEATH` es mayor o menor que 1. Además, al comparar la probabilidad de ‘tiempo de supervivencia mayor que uno’ en los dos grupos, queremos calcular los odds ratio a favor del grupo sin radiación, por lo que este grupo debe tener un código mayor que el otro grupo.

Para conseguir la nueva variable respuesta y la nueva codificación de grupos, vamos a sustituir las variables originales `B3TODEATH` y `GROUP` por las nuevas variables `efectivo` y `grupo`.

	efectivo	grupo	sexo	edad	kps
1	FALSE	Si.Rad	Hombre	>60	75
2	TRUE	Si.Rad	Mujer	<60	50
3	TRUE	Si.Rad	Hombre	>60	90
4	TRUE	Si.Rad	Hombre	>60	100
5	TRUE	Si.Rad	Hombre	>60	95
6	FALSE	Si.Rad	Hombre	>60	80

Table 22: `head(datos)`

	efectivo	grupo	sexo	edad	kps
53	TRUE	No.Rad	Hombre	>60	100
54	TRUE	No.Rad	Hombre	<60	80
55	TRUE	No.Rad	Mujer	<60	100
56	TRUE	No.Rad	Hombre	>60	100
57	FALSE	No.Rad	Hombre	<60	60
58	TRUE	No.Rad	Hombre	>60	100

Table 23: `tail(datos)`

efectivo	grupo	sexo	edad	kps
FALSE:21	Si.Rad:19	Hombre:38	<60:21	Min. : 40.00
TRUE :37	No.Rad:39	Mujer :20	>60:37	1st Qu.: 70.00
				Median : 80.00
				Mean : 80.78
				3rd Qu.: 95.00
				Max. :100.00

Table 24: Sumario

Objetivos del análisis estadístico:

1. Análisis estadístico de los datos. ¿Tiene algún efecto el tratamiento en los resultados? ¿Influyen las covariables?

- Análisis descriptivo

Efectividad del tratamiento por grupos

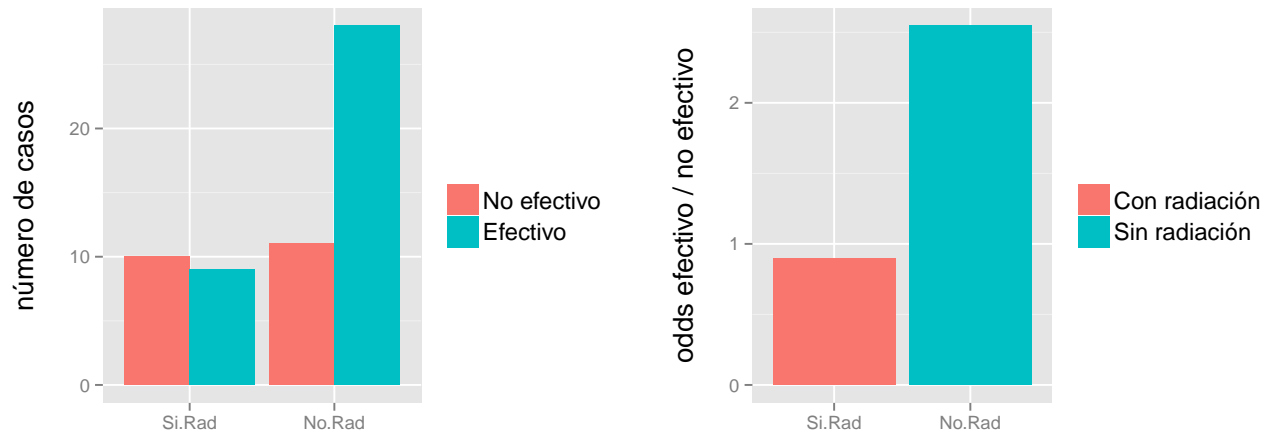


Table 25: Efectividad del tratamiento por grupos

El tratamiento sin radiación parece ser más efectivo que el tratamiento con radiación.

- **Análisis estadístico**

### Comparación de frecuencias observadas-esperadas: Ji-cuadrado

$H_0$  : las variables 'grupo' y 'efectivo' son independientes

$H_1$  : las variables 'grupo' y 'efectivo' no son independientes

	FALSE	TRUE
Si.Rad	10	9
No.Rad	11	28

Table 26: Tabla de contingencia (frecuencias observadas)

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

donde:

$O_i$ : frecuencia observada

$E_i$ : frecuencia esperada (si  $H_0$  es cierta)  $\frac{\text{n}^\circ \text{ casos nivel 'grupo'} * \text{n}^\circ \text{ casos nivel 'efectivo'}}{\text{total casos}}$

$n$ : número de celdas

	FALSE	TRUE
Si.Rad	6.879	12.121
No.Rad	14.121	24.879

Table 27: Tabla de contingencia (frecuencias esperadas)

	statistic	df	p.value
X-squared	3.3	1	0.069

Table 28: Pearson's Chi-squared test

$p - \text{valor} > \alpha = 0.05 \rightarrow$  no se rechazaría  $H_0$  con  $\alpha = 0.05$  (sí con  $\alpha = 0.1$ )

## Comparación de dos proporciones: ARD, RR, OR, NNT

	FALSE	TRUE	
Si.Rad	10	9	19
No.Rad	11	28	39
	21	37	58

	FALSE	TRUE	
Si.Rad	a	b	a+b
No.Rad	c	d	c+d
	a+c	b+d	a+b+c+d

Siendo el riesgo la no efectividad,  $T$  radiación, y  $C$  no radiación:

- $ARD$ : diferencia de riesgos  $\pi_T - \pi_C$ 
  - $AR$ : riesgo absoluto (proporciones)  $\pi_T = \frac{a}{a+b}$ ,  $\pi_C = \frac{c}{c+d}$
- $RR$ : cociente de riesgos  $\frac{\pi_T}{\pi_C}$
- $OR$ : odds ratio  $\frac{O_T}{O_C}$ 
  - $O$ : odds  $\frac{\pi_{no\ ef.}}{\pi_{ef.}} \rightarrow O_T = \frac{a}{b}$ ,  $O_C = \frac{c}{d}$
- $NNT$ : número que se necesita tratar  $\frac{1}{ARD}$

	estimación	IC.95.inf	IC.95.sup	IC.80.inf	IC.80.sup
ARD	0.244	-0.017	0.476	0.070	0.405
RR	1.866	0.967	3.602	1.214	2.869
OR	2.828	0.905	8.835	1.343	5.956
NNT	4.094	NA	NA	2.466	14.289

Table 31: Medidas del efecto del tratamiento - IC 95% y 80%

- $ARD$ : la disminución en el riesgo en el grupo control (no radiación) es de unos 24 pacientes por cada 100 en comparación con el grupo de tratamiento (radiación)
- $RR$ : el cociente de riesgos es  $> 1 \rightarrow$  el grupo de tratamiento presenta mayor riesgo que el grupo control
- $OR$ : el cociente de odds **no efectivo / efectivo**  $> 1 \rightarrow$  el grupo de tratamiento es menos efectivo que el grupo control
- $NNT$ : el número de pacientes que se necesitaría tratar para prevenir una muerte es = 4

## Regresión logística: influencia de covariables

$$\pi_i = \frac{\exp(\beta_0 + \beta_1 X + \dots)}{1 + \exp(\beta_0 + \beta_1 X + \dots)} \rightarrow \text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \log(\text{ODDS}) = \beta_0 + \beta_1 X + \dots$$

$$\text{ODDS} = \exp(\beta_0 + \beta_1 X + \dots)$$

Modelo final:  $\text{logit}(\pi_{\text{efect.}}) = \beta_0 + \beta_1 \text{grupo} + \beta_2 \text{sexo} + \beta_3 \text{edad} + \beta_4 \text{kps} + \beta_2 \beta_3 \text{sexo:edad}$

```
##
## Call:
## glm(formula = efectivo ~ grupo + sexo + edad + kps + sexo:edad,
##      family = binomial(link = logit), data = datos)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.9644  -0.7978   0.4303   0.8109   1.9747
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -6.10097    2.13115  -2.863  0.00420 **
## grupoNo.Rad     1.32609    0.74682   1.776  0.07579 .
## sexoMujer       4.22048    1.60463   2.630  0.00853 **
## edad>60         1.50148    0.89261   1.682  0.09255 .
## kps             0.05606    0.02416   2.320  0.02033 *
## sexoMujer:edad>60 -4.66903    1.81705  -2.570  0.01018 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 75.934  on 57  degrees of freedom
## Residual deviance: 55.347  on 52  degrees of freedom
## AIC: 67.347
##
## Number of Fisher Scoring iterations: 5
```

donde:

- **Intercept:**  $\beta_0 + \beta_1 \text{grupoSi.Rad} + \beta_2 \text{sexoHombre} + \beta_3 \text{edad}<60 + \beta_4 \text{kps}=0 + \beta_2 \beta_3 \text{sexoHombre:edad}<60$
- **grupoNo.Rad:**  $\beta_1 \text{grupoNo.Rad} - \beta_1 \text{grupoSi.Rad}$
- **sexoMujer:**  $\beta_2 \text{sexoMujer} - \beta_2 \text{sexoHombre}$
- **edad>60:**  $\beta_3 \text{edad}>60 - \beta_3 \text{edad}<60$
- **kps:**  $\beta_4 \text{kps}=1 - \beta_4 \text{kps}=0$
- **sexoMujer:edad>60:**  $\beta_2 \beta_3 \text{sexoMujer:edad}>60 - \beta_2 \beta_3 \text{sexoHombre:edad}<60$

$$\rightarrow \exp(\text{grupoNo.Rad}) = \exp(\beta_1 \text{grupoNo.Rad} - \beta_1 \text{grupoSi.Rad}) = \frac{\exp(\beta_1 \text{grupoNo.Rad})}{\exp(\beta_1 \text{grupoSi.Rad})} = \frac{\text{ODDS}_{\text{grupoNo.Rad}}}{\text{ODDS}_{\text{grupoSi.Rad}}}$$

$$\text{ODDS RATIO No.Rad vs Si.Rad} = \exp(\text{grupoNo.Rad})$$

	OR	2.5 %	97.5 %
(Intercept)	0.002	0.000	0.095
grupoNo.Rad	3.766	0.916	18.107
sexoMujer	68.066	4.411	2814.829
edad>60	4.488	0.822	28.938
kps	1.058	1.013	1.116
sexoMujer:edad>60	0.009	0.000	0.242

Table 32: Odds ratios - IC 95%

$\rightarrow$  el cociente de odds efectivo / no efectivo  $> 1 \rightarrow$  el grupo de no radiación es más efectivo que el grupo de radiación (significativo para  $\alpha = 0.1$ )

$$p = \frac{ODDS}{1+ODDS}$$

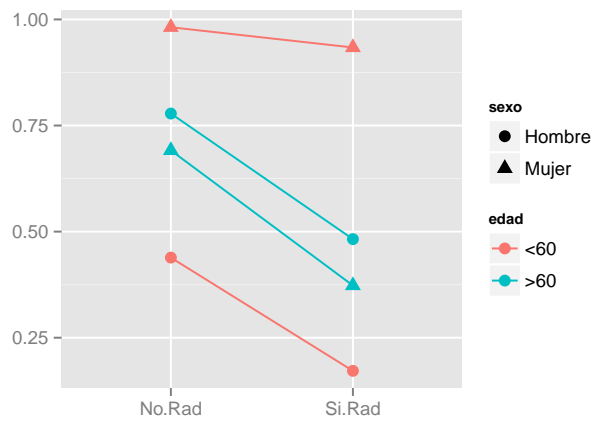
$$\pi_i = \frac{\exp(\beta_0 + \beta_1 \text{grupo} + \beta_2 \text{sexo} + \beta_3 \text{edad} + \beta_4 \text{kps} + \beta_2 \beta_3 \text{sexo:edad})}{1 + \exp(\beta_0 + \beta_1 \text{grupo} + \beta_2 \text{sexo} + \beta_3 \text{edad} + \beta_4 \text{kps} + \beta_2 \beta_3 \text{sexo:edad})} = \frac{\exp(\text{predictor lineal})}{1 + \exp(\text{predictor lineal})}$$

grupo	sexo	edad	kps	prob
No.Rad	Hombre	<60	80.776	0.439
No.Rad	Hombre	>60	80.776	0.778
No.Rad	Mujer	<60	80.776	0.982
No.Rad	Mujer	>60	80.776	0.691
Si.Rad	Hombre	<60	80.776	0.172
Si.Rad	Hombre	>60	80.776	0.482
Si.Rad	Mujer	<60	80.776	0.934
Si.Rad	Mujer	>60	80.776	0.373

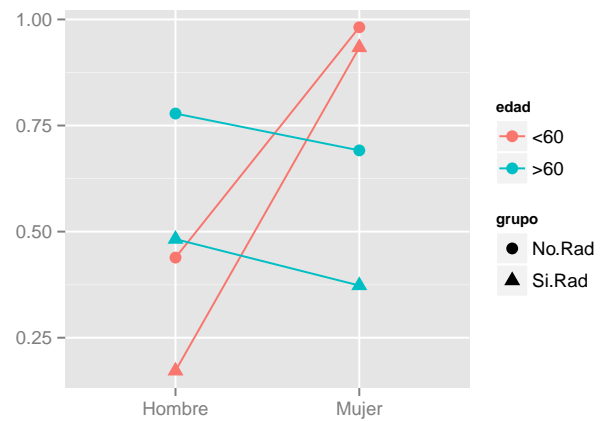
Table 33: Probabilidades estimadas de efectividad para KPS medio



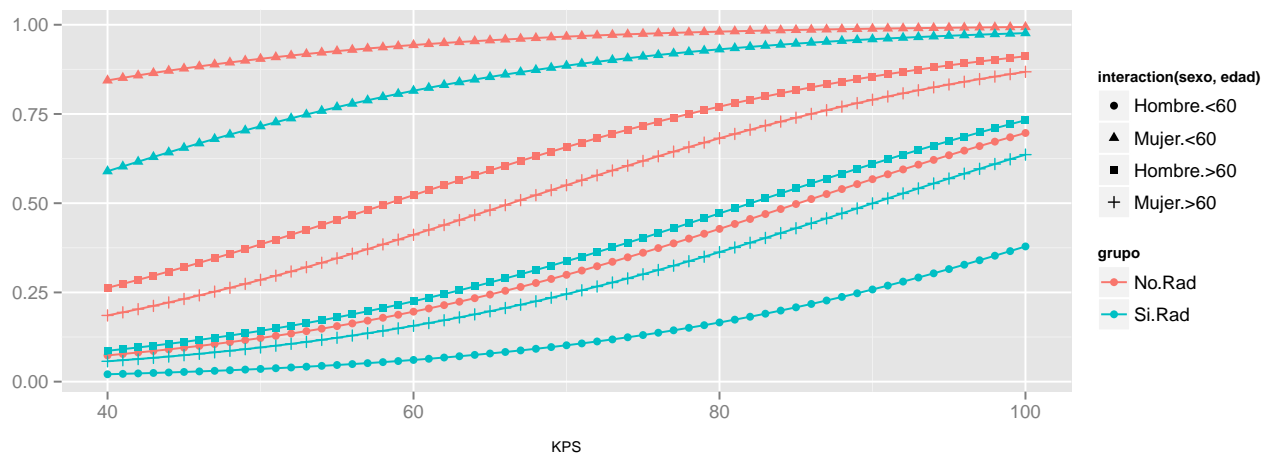
Probabilidades estimadas de efectividad de los tratamientos  
(para KPS medio)



Efecto de la interaccion sexo–edad sobre la efectividad  
(para KPS medio)



Probabilidades estimadas de efectividad



2. Se va a realizar un ensayo algo más grande en el que se va a utilizar un nivel de significatividad  $\alpha = 0.05$ . Se desea obtener una potencia del 80% para detectar probabilidades  $p_A = 0.5$  y  $p_B = 0.7$ . ¿Cuántos pacientes por grupo se necesitan? ¿Variaría mucho ese tamaño si se quisiera una potencia del 90% o del 95%? ¿Y si  $p_B = 0.6$ ?

## **Informe final**

blablabla

## ANEXO II

```
sessionInfo()
```

```
## R version 3.1.2 (2014-10-31)
## Platform: x86_64-pc-linux-gnu (64-bit)
##
## locale:
##  [1] LC_CTYPE=en_GB.UTF-8      LC_NUMERIC=C
##  [3] LC_TIME=en_GB.UTF-8      LC_COLLATE=en_GB.UTF-8
##  [5] LC_MONETARY=en_GB.UTF-8  LC_MESSAGES=en_GB.UTF-8
##  [7] LC_PAPER=en_GB.UTF-8     LC_NAME=C
##  [9] LC_ADDRESS=C             LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_GB.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] grid      splines  stats    graphics grDevices utils      datasets
## [8] methods  base
##
## other attached packages:
## [1] gridExtra_0.9.1 reshape2_1.4    doBy_4.5-11    MASS_7.3-35
## [5] survival_2.37-7 knitr_1.7      ggplot2_1.0.0
##
## loaded via a namespace (and not attached):
##  [1] colorspace_1.2-4 digest_0.6.3    evaluate_0.5.5  formatR_1.0
##  [5] gtable_0.1.2    htmltools_0.2.6 labeling_0.3     lattice_0.20-29
##  [9] Matrix_1.1-4    munsell_0.4.2  plyr_1.8.1      proto_0.3-10
## [13] Rcpp_0.11.3     rmarkdown_0.3.12 scales_0.2.4    stringr_0.6.2
## [17] tools_3.1.2     yaml_2.1.13
```