



Introducción al manejo del software estadístico R



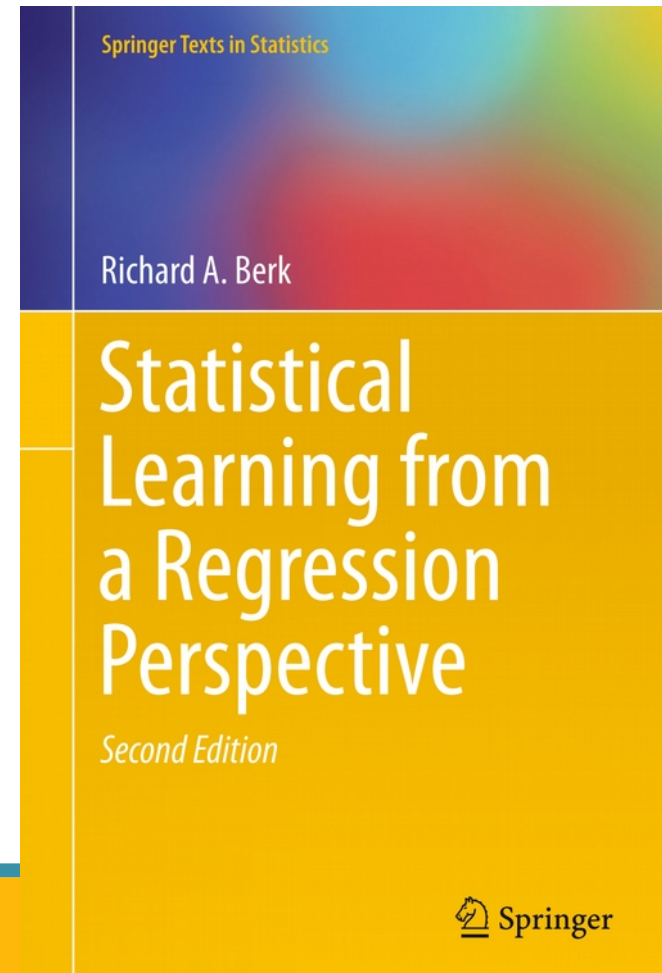
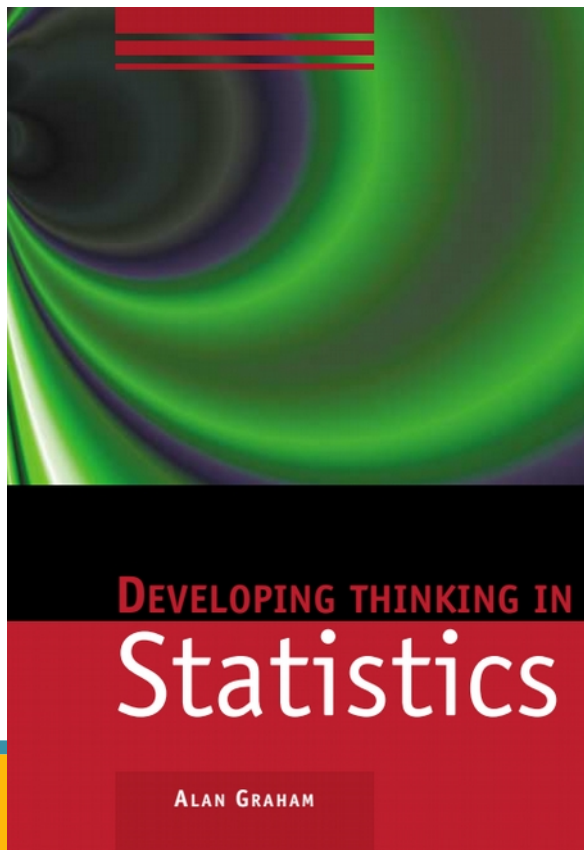
Jorge Jiménez
Escuela de Biología, USAC

Contenido / objetivos

- Referencias útiles
- Conceptos generales (2)
- Procedimiento 1
 - Importar datos, gráficas y estadísticos descriptivos, pruebas de hipótesis
- Procedimiento 2
 - Repaso

Referencias útiles

- ◆ Desarrollo del pensamiento estadístico
- ◆ Enfoque de causalidad



Conceptos generales 1

Lenguaje de programación en R:

- **Objetos** = sustantivos
- **Funciones** = verbos

Podemos construir oraciones:

```
summary(datos1$sex)
```

```
summary(datos1[, 5])
```

Procedimiento 1

¿Existe discriminación sexual en los salarios de los profesores de la Universidad X?

Pasos 1 y 2

Paso 1: camino de hormigas

Paso 2:

```
read.table("clipboard") -> datos1
```

Paso 3

Paso 3:

```
sapply(datos1, "class")
```

```
> sapply(datos1, "class")  
      rank discipline      yrs1      yrs2  
"factor"  "factor"  "integer"  "integer"  "fa
```

Factor = categórica; Numeric = numérica, continua

Integer = numérica, discreta

Paso 4

Paso 4:

```
summary(datos1$salary)
```

```
sd(datos1$salary)
```

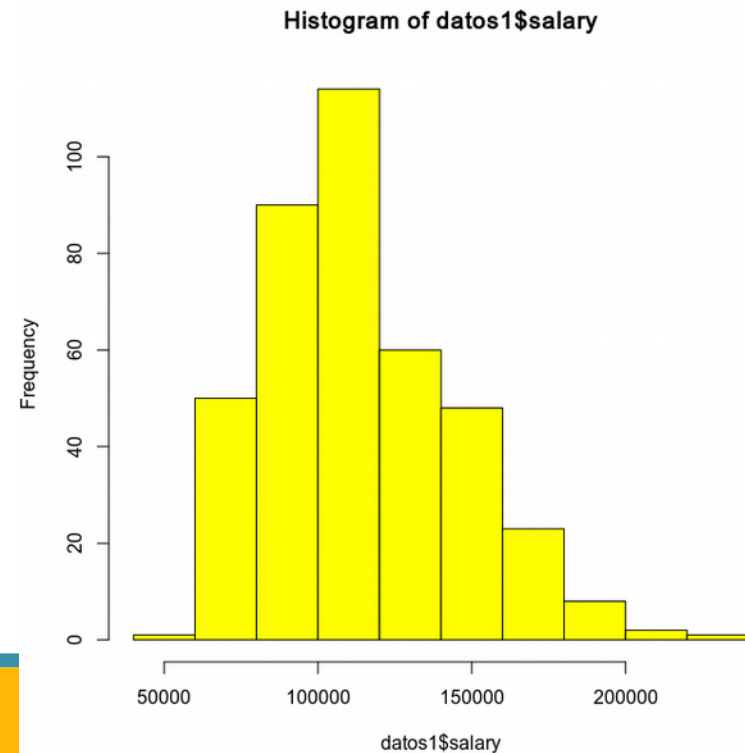
```
> summary(datos1$salary)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 57800   91000  107300  113706  134185  231545

> sd(datos1$salary)
[1] 30289.04
```

Paso 5

Paso 5:

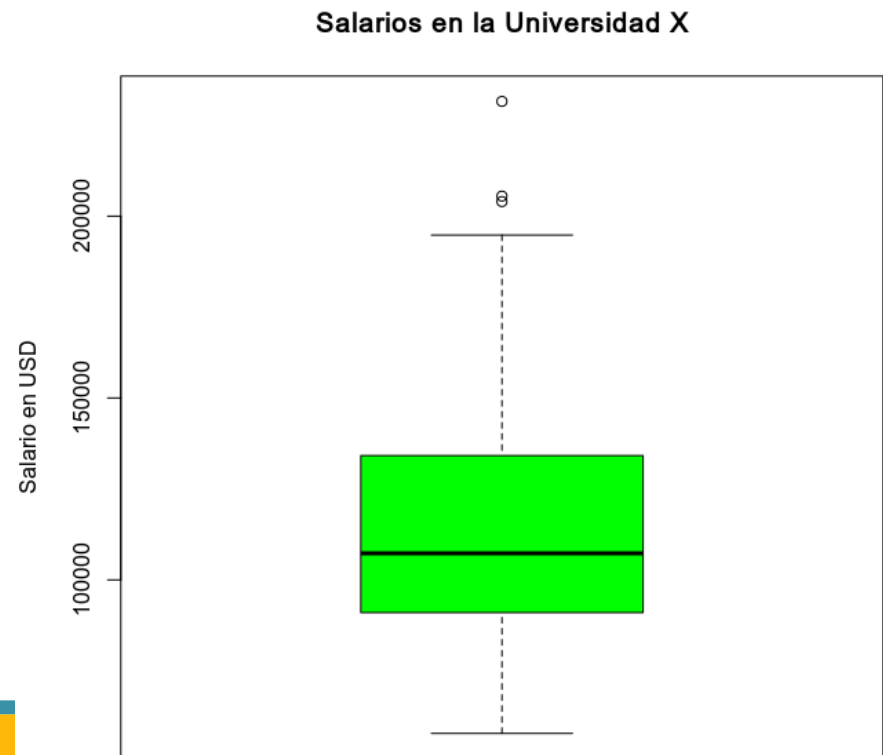
```
hist(datos1$salary, col="yellow")
```



Paso 6

```
boxplot(datos1$salary, col="green")
```

```
> boxplot(datos1$salary,col="green",ylab="Salario en USD",main="Salarios en la Universidad X")
```



Paso 7

```
shapiro.test(datos1$salary)
```

```
> shapiro.test(datos1$salary)
```

```
      Shapiro-Wilk normality test
```

```
data:  datos1$salary
```

```
W = 0.95988, p-value = 6.076e-09
```

Conceptos generales 2

La ecuación de la recta

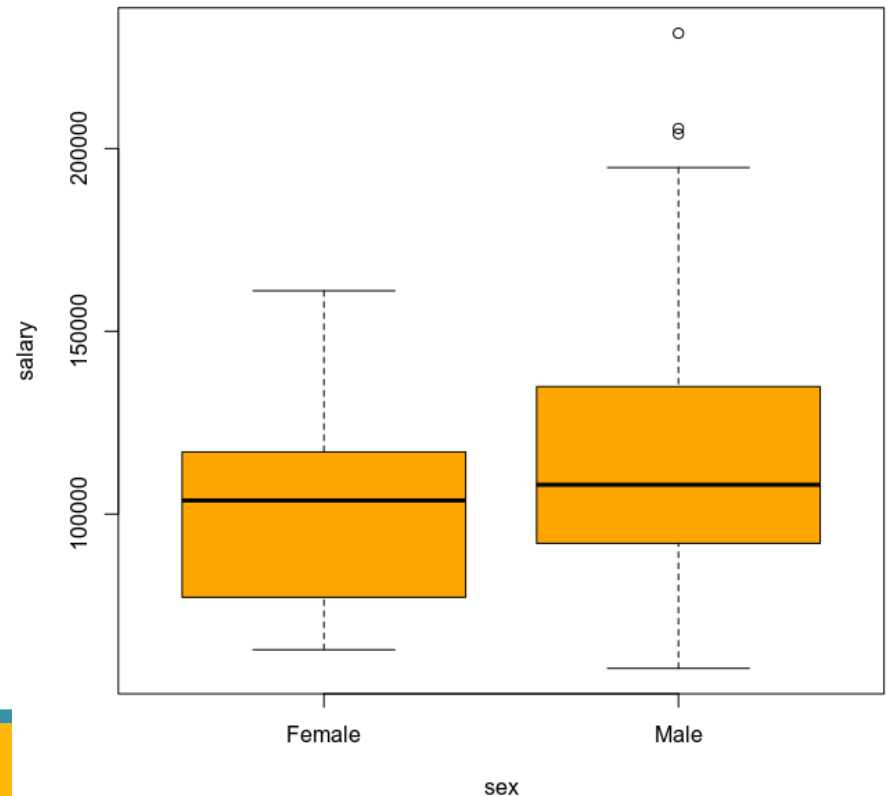
- $y = ax + b$
- Respuesta ~ Explicativas
- Efectos ~ Causas

El argumento fórmula

```
lm(salary~sex, data=datos1)
```

Paso 8

```
boxplot(salary~sex, data=datos  
1, col="orange")
```



Paso 9

¿Existe diferencia?

```
t.test(salary~sex, data=datos1)
```

```
> t.test(salary~sex, data=datos1)
```

Welch Two Sample t-test

```
data: salary by sex
t = -3.1615, df = 50.122, p-value = 0.002664
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -23037.916  -5138.102
sample estimates:
mean in group Female    mean in group Male
      101002.4           115090.4
```

Paso 10

¿Es menor en mujeres?

```
t.test(salary~sex, data=datos1,  
alternative="less")
```

```
> t.test(salary~sex, data=datos1, alternative="less")
```

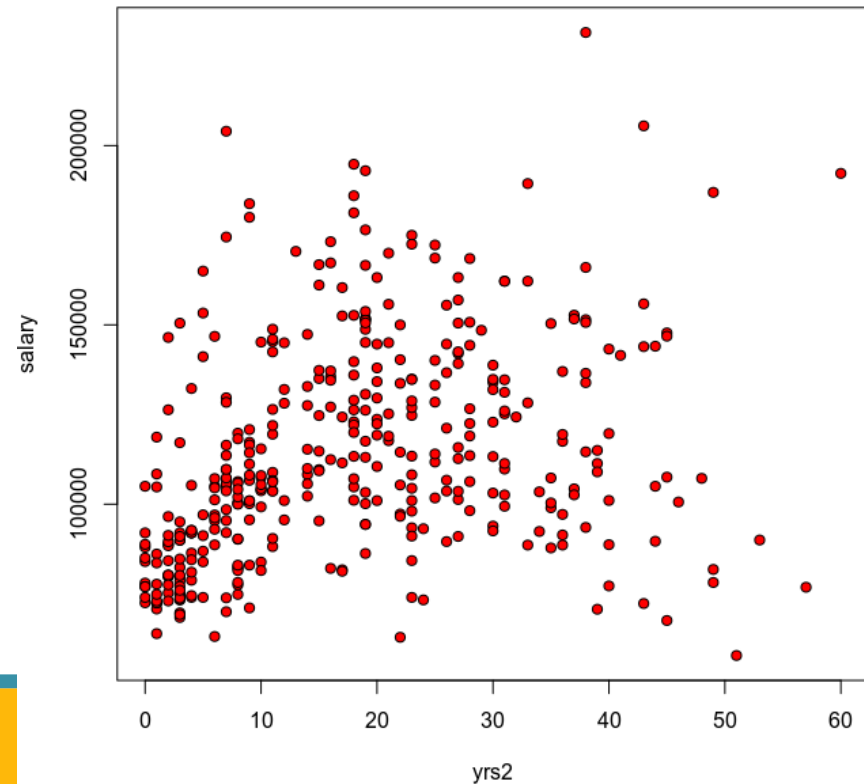
Welch Two Sample t-test

```
data: salary by sex  
t = -3.1615, df = 50.122, p-value = 0.001332  
alternative hypothesis: true difference in means is less than 0  
95 percent confidence interval:  
-Inf -6620.263  
sample estimates:  
mean in group Female    mean in group Male  
101002.4                115090.4
```

Paso 11

¿Se relaciona con los años de servicio?

```
plot(salary~yrs2, data=datos1,  
pch=21, bg="red")
```



Paso 12

¿Se relaciona con los años de servicio?

```
cor.test(datos1$salary, datos1$yrs2, method="pearson")
```

```
> cor.test(datos1$salary, datos1$yrs2, method="pearson")
```

Pearson's product-moment correlation

data: datos1\$salary and datos1\$yrs2

t = 7.0602, df = 395, p-value = 7.529e-12

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.2443740 0.4193506

sample estimates:

cor

0.3347447

Paso 13

¿Cuáles son las variables importantes?

```
summary(lm(salary~., data=datos1))
```

Residuals:

Min	1Q	Median	3Q	Max
-65248	-13211	-1775	10384	99592

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	78862.8	4990.3	15.803	< 2e-16	***
rankAsstProf	-12907.6	4145.3	-3.114	0.00198	**
rankProf	32158.4	3540.6	9.083	< 2e-16	***
disciplineB	14417.6	2342.9	6.154	1.88e-09	***
yrs1	535.1	241.0	2.220	0.02698	*
yrs2	-489.5	211.9	-2.310	0.02143	*
sexMale	4783.5	3858.7	1.240	0.21584	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 22540 on 390 degrees of freedom

Multiple R-squared: 0.4547, Adjusted R-squared: 0.4463

F-statistic: 54.2 on 6 and 390 DF, p-value: < 2.2e-16

Paso 14

¿Cuál podría ser el mejor modelo?

```
summary(lm(salary~rank, data=datos1))
```

Residuals:

Min	1Q	Median	3Q	Max
-68972	-16376	-1580	11755	104773

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	93876	2954	31.777	< 2e-16	***
rankAsstProf	-13100	4131	-3.171	0.00164	**
rankProf	32896	3290	9.997	< 2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 23630 on 394 degrees of freedom

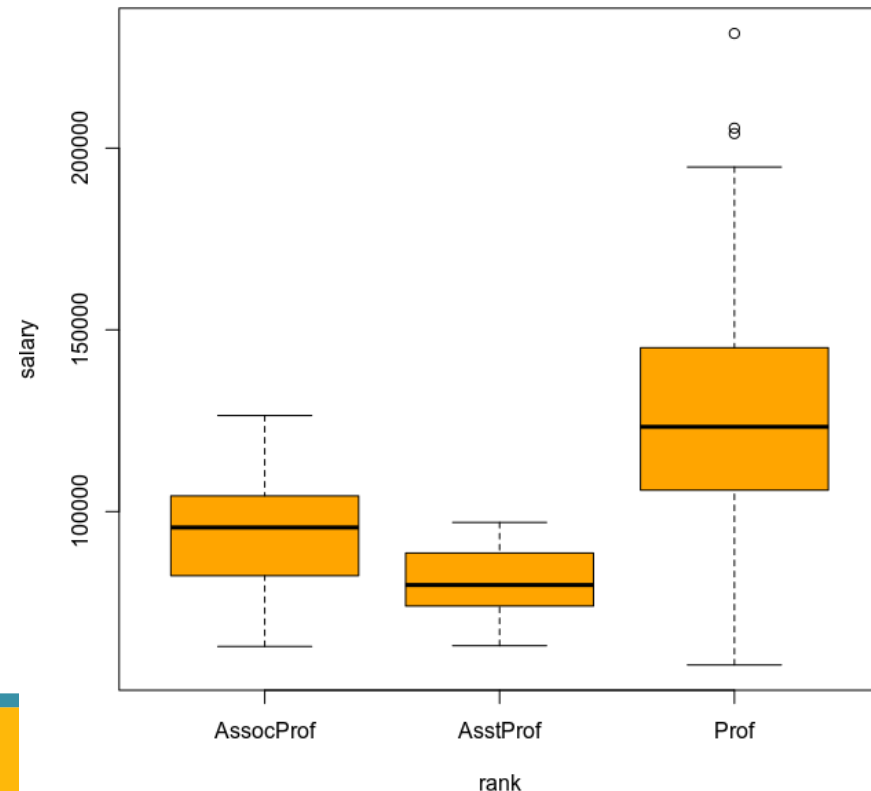
Multiple R-squared: 0.3943, Adjusted R-squared: 0.3912

F-statistic: 128.2 on 2 and 394 DF, p-value: < 2.2e-16

Paso 15

¿Es tan importante el cargo?

```
boxplot(salary~rank, data=datos1,  
        col="orange")
```



Paso 16a

¿Cómo se relaciona con el cargo?

```
anova(lm(salary~rank, data=datos1))
```

```
> anova(lm(salary~rank, data=datos1))
```

```
Analysis of Variance Table
```

```
Response: salary
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
rank	2	1.4323e+11	7.1616e+10	128.22	< 2.2e-16 ***
Residuals	394	2.2007e+11	5.5855e+08		

Paso 16b

¿Cuál es el cargo con salario distinto?

```
pairwise.t.test(datos1$salary,  
datos1$rank,p.adj="bonf")
```

Pairwise comparisons using t tests with pooled SD

data: datos1\$salary and datos1\$rank

	AssocProf	AsstProf
AsstProf	0.0049	-
Prof	<2e-16	<2e-16

P value adjustment method: bonferroni

Descanso intermedio



Procedimiento 2

¿Existe discriminación sexual en los salarios de los trabajadores de la Empresa X?

Gracias
Jorge Jiménez
Escuela de Biología, USAC
jimenez.jorge@usac.edu.gt

Pteris muricella Fée

