PRÁCTICA 2 ESTADÍSTICA APLICADA

1. A partir del *data frame* CHFLS utilizado en clase, genera un nuevo vector que valga 0 (o *FALSE*) si el nivel educativo de la respondente (*R_edu*) es menor que bachillerato completo (*Senior High School*) y 1 (o *TRUE*) en otro caso. Emplea este vector para generar un nuevo vector de tipo factor denominado *R_edu2* con dos niveles (utilizando los valores del vector anterior), da a cada nivel un nombre adecuado a lo que representa y crea un nuevo *data frame* añadiendo este vector a CHFLS.

	R_region	R_age	R_edu	R_income	R_health	R_height	R_happy	A_height	A_edu	A_income
2	Northeast	54	Senior high school	900	Good	165	Somewhat happy	172	Senior high school	500
3	Northeast	46	Senior high school	500	Fair	156	Somewhat happy	170	Senior high school	800
10	Northeast	48	Senior high school	800	Good	163	Somewhat happy	172	Junior high school	700
11	Northeast	46	Junior high school	300	Fair	164	Somewhat happy	174	Elementary school	700

Así es como se ve el fichero CHFLS sin modificar.

	R_region	R_age	R_edu	R_income	R_health	R_height	R_happy	A_height	A_edu	A_income	R_edu2
2	Northeast	54	Senior high school	900	Good	165	Somewhat happy	172	Senior high school	500	CON BACH
3	Northeast	46	Senior high school	500	Fair	156	Somewhat happy	170	Senior high school	800	CON BACH
10	Northeast	48	Senior high school	800	Good	163	Somewhat happy	172	Junior high school	700	CON BACH
11	Northeast	46	Junior high school	300	Fair	164	Somewhat happy	174	Elementary school	700	SIN BACH

Hemos decidido llamar “SIN BACH” y “CON BACH” a los dos niveles educativos, así se ve el nuevo *data frame* una vez añadida la columna “*R_edu2*”.

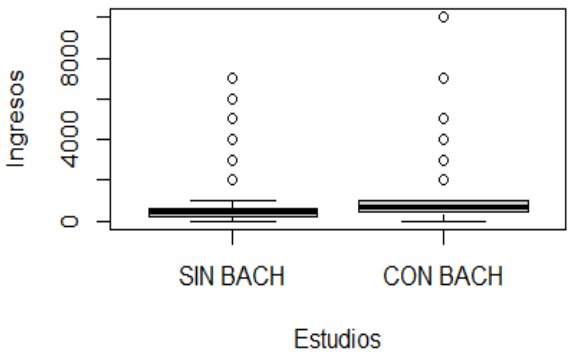
2. Estudia mediante herramientas gráficas la relación entre las variables *R_income* y *R_edu2*. ¿Parece haber diferencias relevantes de salario en función del nivel educativo? Emplea a continuación herramientas inferenciales para estudiar si existen diferencias significativas en los salarios para ambos niveles educativos. Interpreta el resultado. ¿Qué diferencia existe entre lo que permiten concluir las herramientas gráficas y las inferenciales?

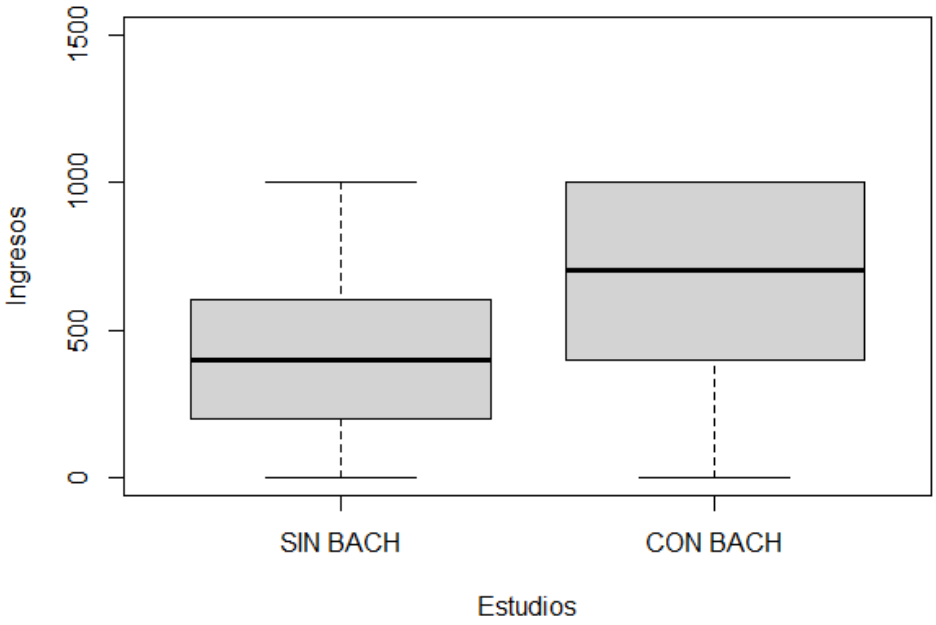
Herramientas gráficas

Hemos realizado un diagrama de caja y bigotes (boxplot) para hacernos una idea general.

A simple vista parece que las personas con estudios tienen mayores ingresos de media, para verlo mejor y sacar conclusiones, haremos un “zoom” a la caja y bigotes.

En cuanto a los datos atípicos, sí que se observan mayores sueldos en las personas con estudios superiores.





Reduciendo el límite superior de ingresos a 1500, podemos observar de manera clara las cajas de ambos niveles.

En cuanto al grupo que no finalizó el bachillerato completo, los sueldos típicos van de unos 250 a poco más de 500. Aquellos que sí lo acabaron ingresan desde algo menos de 500 hasta 1000, lo que supone prácticamente el doble de ingresos de media.

Herramientas inferenciales

Usamos el shapiro test para ver si se distribuyen los datos como una normal.

El test rechaza que los datos se distribuyan mediante una normal, para ambos casos.

Por tanto, usamos un contraste de hipótesis no paramétrico, con el que podremos ver si hay una fuerte relación o no, entre los salarios y sus niveles educativos.

```
> tapply(CHFLS2$R_income, CHFLS2$R_edu2, shapiro.test)
$`SIN BACH`

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.56881, p-value < 2.2e-16


$`CON BACH`

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.60345, p-value < 2.2e-16
```

La hipótesis nula es tener todas las medias iguales.

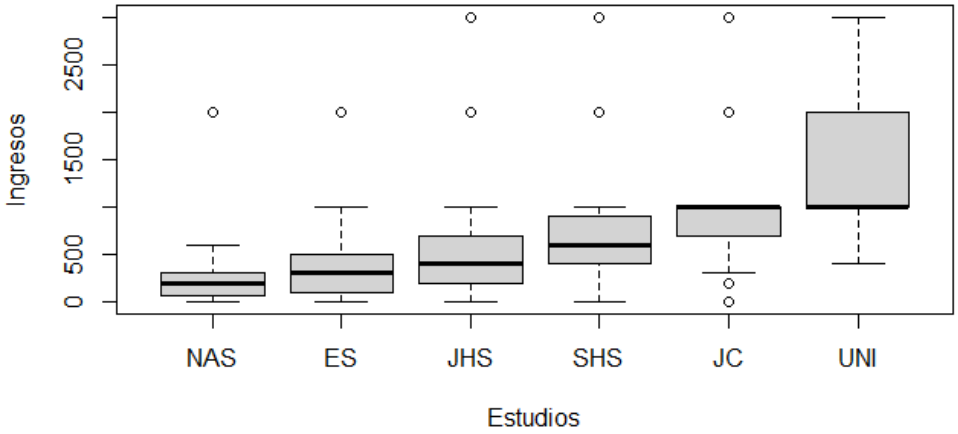
Siendo un p-valor tan pequeño y menor a 0.05, rechazamos Ho, lo que nos permite concluir que sí existen diferencias relevantes de salario en función del nivel educativo.

```
> wilcox.test(R_income~R_edu2, data=CHFLS, conf.int=TRUE)

        Wilcoxon rank sum test with continuity correction

data:  R_income by R_edu2
W = 158060, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
 -340 -300
sample estimates:
difference in location
                  -300
```

3. Realiza el mismo estudio del apartado 2 pero ahora empleando el factor original *R_edu* en lugar de *R_edu2*. ¿Qué se observa mediante las herramientas gráficas? ¿Qué herramienta inferencial se tendría que emplear ahora para estudiar si hay diferencias de salario entre los 6 niveles de *R_edu*?

Herramientas gráficas

Comenzamos de nuevo haciendo un diagrama de caja y bigotes.



Este diagrama nos permite ver de forma muy clara que el nivel de ingresos tiende a aumentar a medida que aumentan los estudios. El único grupo que supera un salario de 1000 con asiduidad es el de “University”, siendo este el más amplio tanto en su RIC (caja), como en su rango (bigotes). Por otro lado, en los dos primeros grupos, raro es el caso en el que se superan la franja de los 500.

También es curioso el caso de las medianas: se encuentran bastante centradas en las cajas de los 4 primeros grupos mientras que en los casos de “Junior College” y “University” están en los extremos. Para “Junior College” coincide prácticamente con su cuartil superior, cerca de 1000, lo que significa que un alto porcentaje de personas de este grupo tiene un mismo salario. Para “University” ocurre algo similar, la mediana coincide con el cuartil inferior y, a su vez, con la mediana y cuartil superior del grupo anterior.

Herramientas inferenciales

Queremos estudiar la diferencia de salarios entre los 6 niveles de educación. Para ello lo mejor es utilizar una tabla ANOVA, con el fin de contrastar la hipótesis nula, donde las medias de los grupos son iguales, frente a que al menos 2 de ellas difieran.

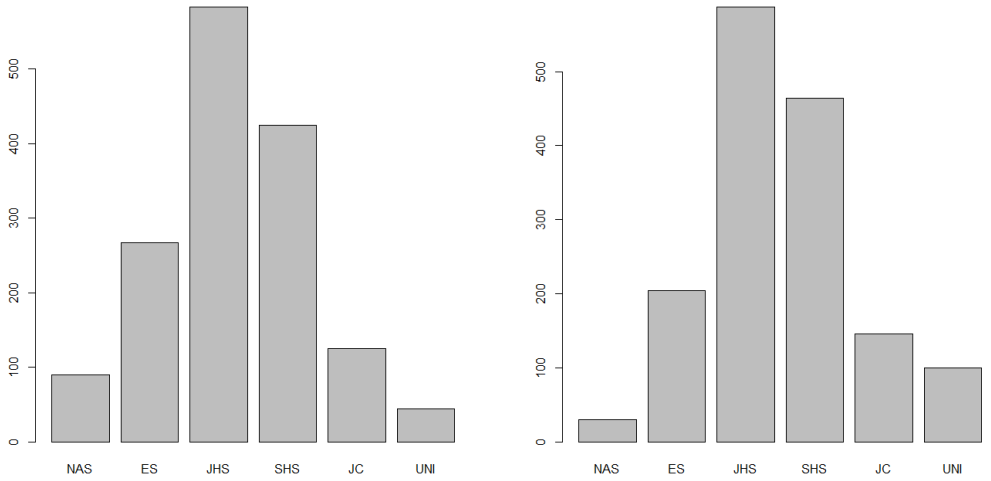
El problema es que los grupos con los que queremos trabajar no cuentan con el mismo número de observaciones cada uno, por lo que no podemos realizar la tabla ANOVA directamente. En su lugar, hemos decidido utilizar una tabla de contingencia:

	R_income																							
R_edu	0	10	20	30	40	50	60	70	80	90	100	200	300	400	500	600	700	800	900	1000	2000	3000	4000	5000
Never attended school	8	1	2	4	5	0	3	0	3	0	17	21	9	8	5	3	0	0	0	0	1	0	0	0
Elementary school	26	2	2	2	4	1	7	1	9	1	21	36	39	34	28	17	10	6	1	11	7	0	0	0
Junior high school	68	1	0	0	3	5	1	2	4	2	24	51	63	82	69	61	36	34	12	37	19	4	2	1
Senior high school	27	0	0	3	1	2	0	0	2	1	8	18	37	52	56	40	36	30	16	67	25	3	0	1
Junior college	6	0	0	0	0	0	0	0	0	0	0	2	2	7	7	6	9	12	5	45	16	6	0	0
University	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	3	2	2	1	15	13	4	1	1

En esta tabla se observa que el desempleo es mayor en grupos con estudios elementales. De igual manera, es claro que cada grupo se distribuye y concentra en franjas de mayores salarios a medida que aumentan sus estudios.

4. Estudia mediante herramientas gráficas la relación entre los factores *R_edu* y *A_edu*. ¿Se observa una dependencia entre ambas variables? Interpreta el resultado.

Comenzamos estudiando ambas distribuciones con un diagrama de barras para hacernos una idea general.

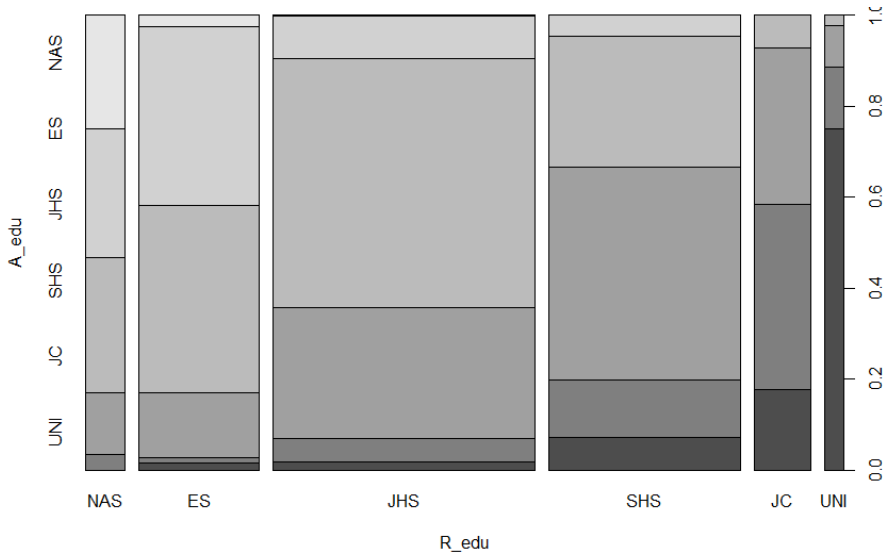


El diagrama de la izquierda muestra la distribución de los estudios de las mujeres y el de la derecha el de sus parejas.

A simple vista se aprecia una alta relación. Usemos ahora un gráfico de columnas, útil para 2 variables categóricas. Pero previamente veamos la tabla de contingencia.

A_edu \ R_edu	Never attended school	Elementary school	Junior high school	Senior high school	Junior college	University
Never attended school	22	7	1	0	0	0
Elementary school	25	104	55	20	0	0
Junior high school	26	110	319	122	9	1
Senior high school	12	38	168	199	43	4
Junior college	3	3	29	54	51	6
University	0	4	11	30	22	33

La tabla de contingencia es muy ilustradora: las parejas tienen mayoritariamente un nivel de estudios muy similar.



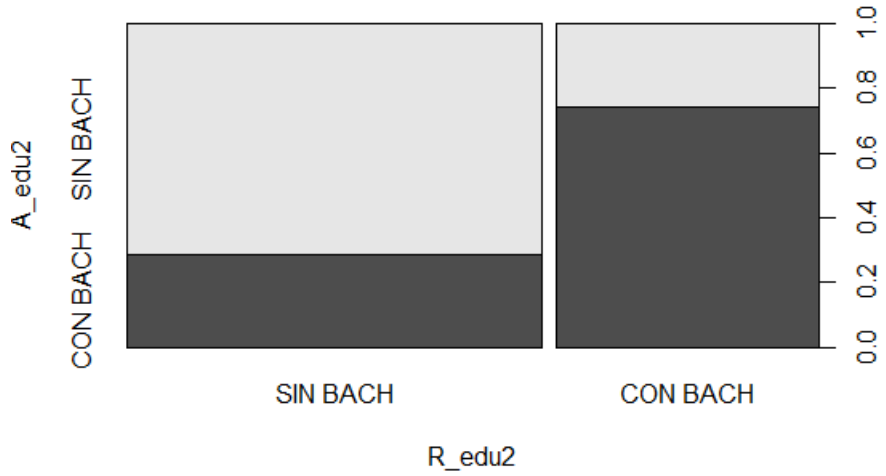
El gráfico de columnas viene a representar gráficamente lo que hemos inferido al mirar la tabla, los niveles de estudio de las mujeres se concentran en torno al nivel de estudio de sus parejas.

5. Realiza con *A_edu* una recodificación similar a la del apartado 1 para generar un nuevo factor binario *A_edu2*. Estudia ahora, mediante herramientas gráficas e inferenciales, la relación entre los factores *A_edu2* y *R_edu2*. Interpreta el resultado.

Herramientas gráficas

Dado que estamos trabajando con 2 variables categóricas, volvemos a recurrir al gráfico de columnas y a su tabla de contingencia.

R_edu2 \ A_edu2	SIN BACH	CON BACH
SIN BACH	669	268
CON BACH	152	442



Al tener tan solo dos grandes grupos, se ve perfectamente que las personas que acabaron el bachillerato suelen tener parejas que también lo hicieron; de la misma forma, aquellos que no lo acabaron, suelen tener parejas que tampoco lo hicieron.

Herramientas inferenciales

Utilizamos el contraste de independencia de McNemar's por tratarse de variables categóricas pareadas.

```
        McNemar's Chi-squared test

data:  tab_edu
McNemar's chi-squared = 32.038, df = 1, p-value = 1.512e-08
```

El p-valor es muy pequeño (menor a 0.05). Por tanto, la evidencia para rechazar la independencia de ambas variables (Ho) es muy elevada. En otras palabras, las variables son dependientes.