Probabilidad y Estadística Fundamental

Resumen y descripción de datos de dos variables

Profesor: Nicolás López

Universidad Nacional de Colombia

Contenido

Contenido

Introducción

Cuando dos variables se miden en una sola unidad experimental, los datos resultantes se denominan datos bivariados. De estos, surgen las siguientes preguntas:

- ¿Cómo se deben presentar estos datos?
- Cada variable tiene determinadas características individuales ¿cómo estudio la **relación** entre las variables?

Contenido

Datos cualitativos-cualitativos

Además de la variable *escolaridad* para la muestra de $n = 3000$ personas, se cuenta con la variable *raza*. ¿Cómo se distribuye esta nueva variable? ¿cómo se relaciona con la variable *escolaridad*?

Datos cualitativos-cualitativos

Distribuciones **marginales**. Gráfico de barras

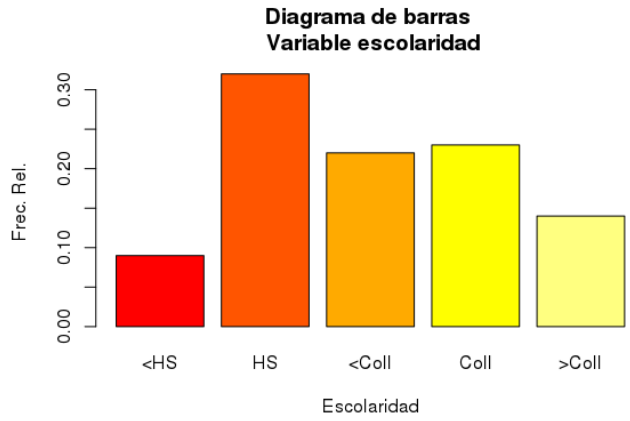


Figura 1: Diagrama de barras para la variable *Escolaridad*. Frecuencia relativa.

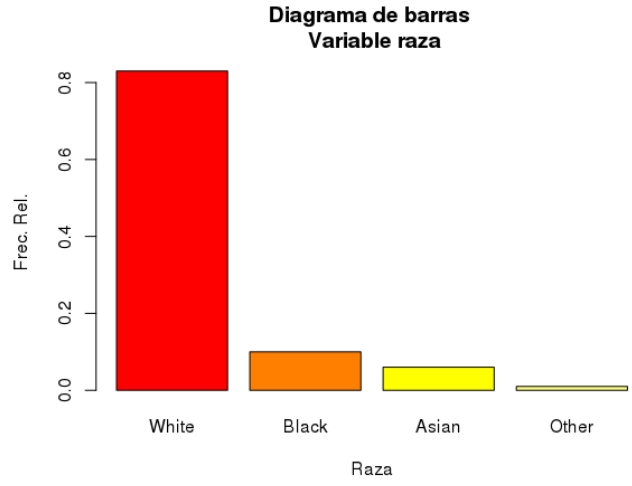
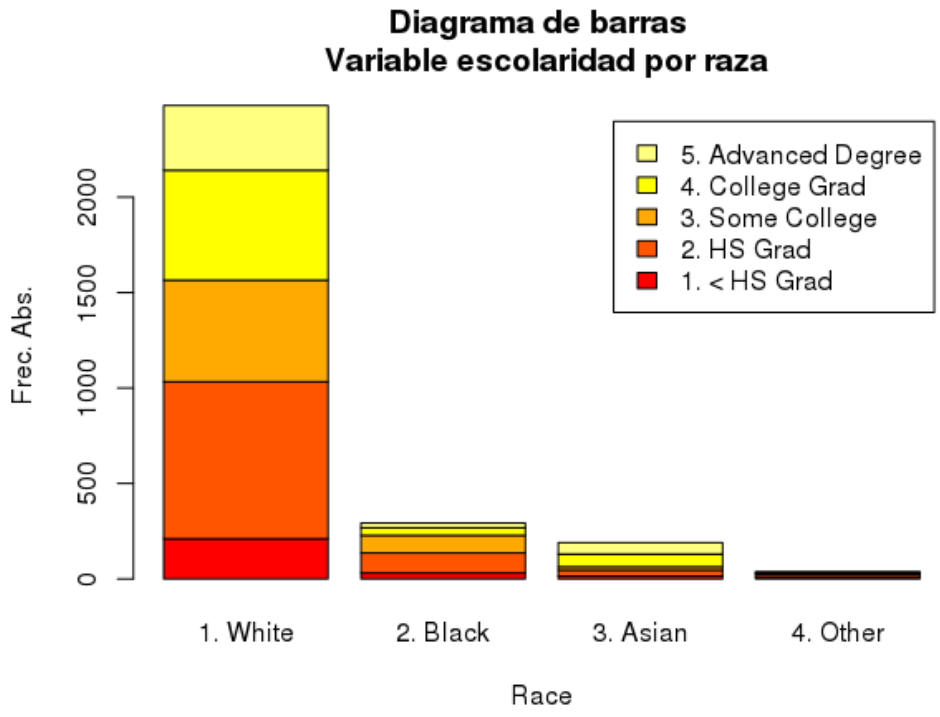


Figura 2: Diagrama de barras para la variable *Raza*. Frecuencia relativa.

Datos cualitativos-cualitativos

Distribuciones **conjuntas**. Gráficas de barras apiladas.

Figura 3: Diagrama de barras para la variable *escolaridad* en los niveles de la variable *raza*. Frecuencia absoluta.



Datos cualitativos-cualitativos

Distribuciones **conjuntas** y **condicionadas**. Gráficas de barras lado a lado.

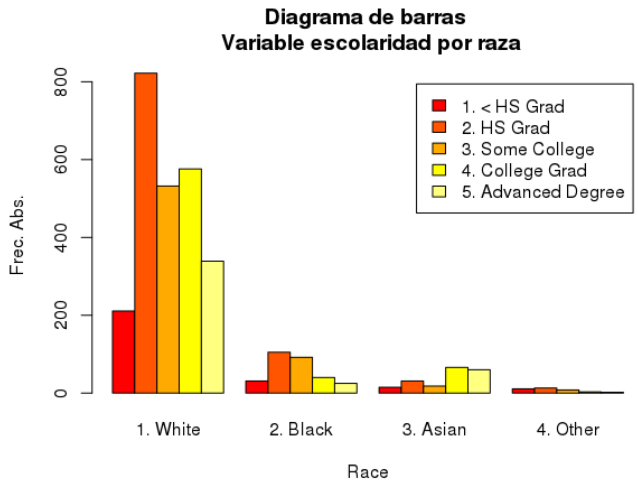


Figura 4: Diagrama de barras para la variable *escolaridad* en los niveles de la variable *raza*. Frecuencia absoluta.

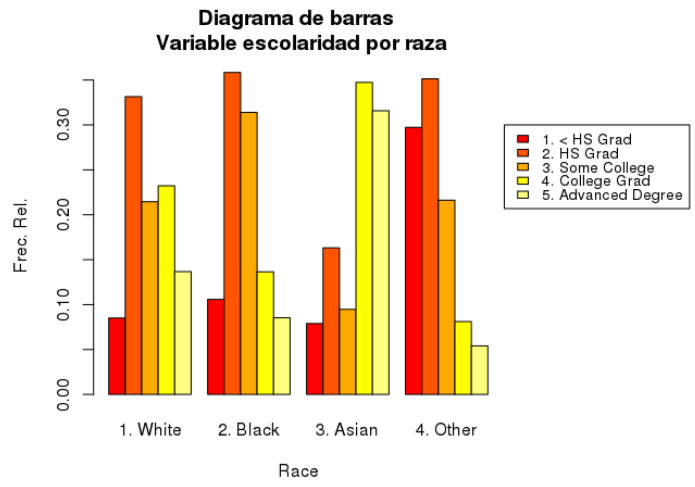
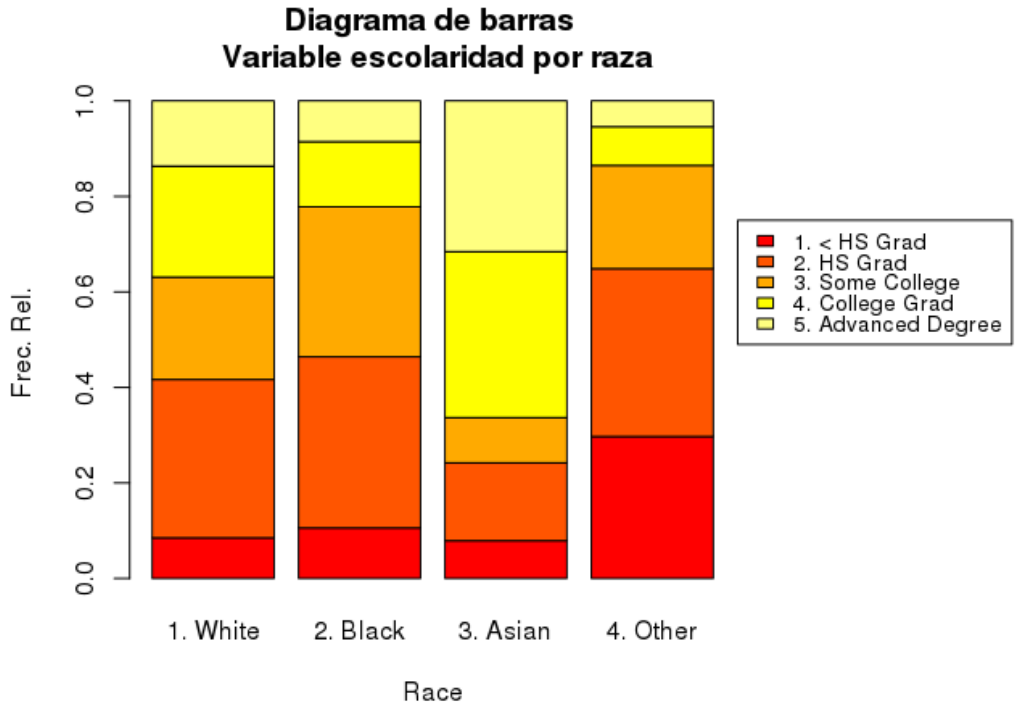


Figura 5: Diagrama de barras para la variable *escolaridad* **condicionado** a los niveles de la variable *raza*. Frecuencia relativa.

Datos cualitativos-cualitativos

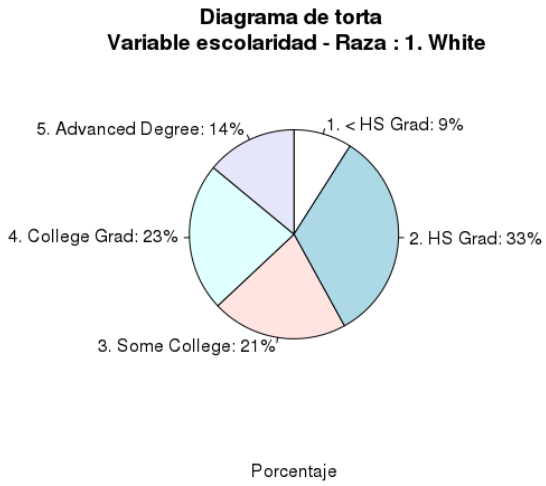
Distribuciones **condicionadas**. Gráficas de barras apiladas

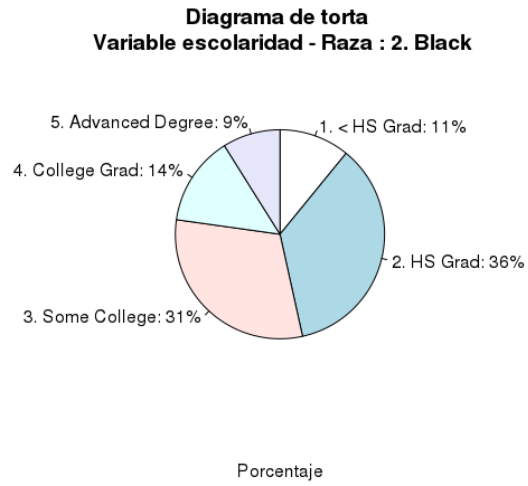
Figura 6: Diagrama de barras para la variable *escolaridad* condicionado a los niveles de la variable *raza*. Frecuencia relativa.

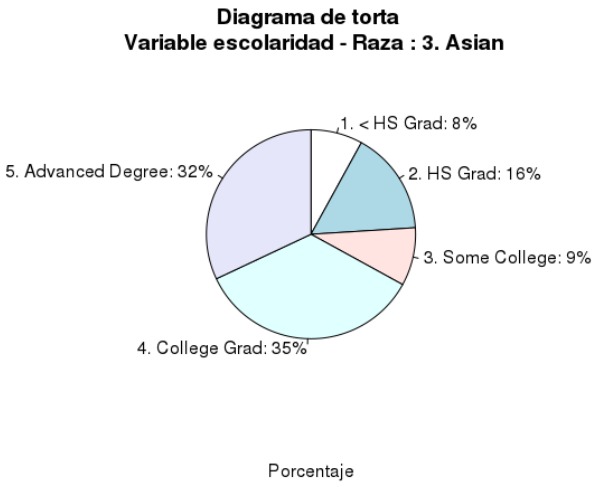


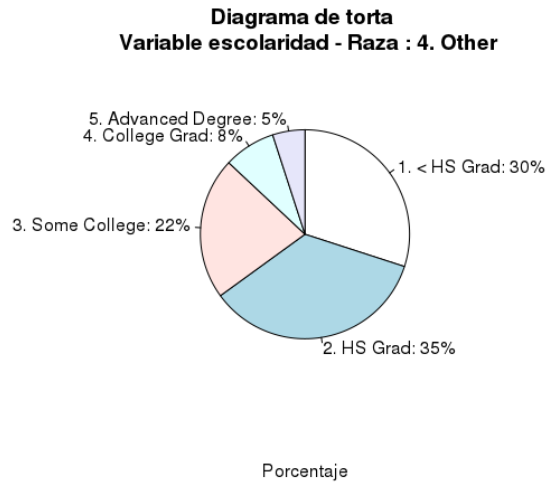
Datos cualitativos-cualitativos

Distribuciones **condicionadas**. Gráficas de pastel lado a lado









Datos cualitativos-cualitativos

Representación tabular

Notas

- La distribuciones marginales no permiten ver la asociación entre las variables.
- Se distingue entre frecuencias relativas conjunta y condicionada para la realización de los gráficos.
- Para la distribución conjunta, Los gráficos de barras lado a lado pueden ser más claros que los apilados.
- Para la distribución condicionada, Tanto el gráfico de barras apilado como el lado a lado presentan la distribución de manera eficiente. El gráfico de pastel agrupado es evidentemente menos adecuado.

Datos cualitativos-cualitativos

Representación tabular

Se puede explorar la distribución bivariada mediante la representación tabular del mismo conjunto de datos

	1. White	2. Black	3. Asian	4. Other	Suma
1. < HS Grad	211	31	15	11	268
2. HS Grad	822	105	31	13	971
3. Some College	532	92	18	8	650
4. College Grad	576	40	66	3	685
5. Advanced Degree	339	25	60	2	426
Suma	2480	293	190	37	3000

Cuadro 1: Tabla de contingencia bidimensional. Educación por raza.

Datos cualitativos-cualitativos

Representación tabular. Tabla de contingencia bidimensional

Una **tabla de contingencia bidimensional** permite clasificar a las unidades estadísticas en categorías *exhaustivas* y *mutuamente excluyentes* de dos variables cualitativas.

K	1	...	j	...	p	Suma
1	k_{11}	...	k_{1j}	...	k_{1p}	$k_{1.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	k_{i1}	...	k_{ij}	...	k_{ip}	$k_{i.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	k_{n1}	...	k_{nj}	...	k_{np}	$k_{n.}$
Suma	$k_{,1}$	...	$k_{.j}$	...	$k_{.p}$	$n = k_{..}$

Cuadro 2: Tabla de contingencia bidimensional. Forma general.

Datos cualitativos-cualitativos

Representación tabular

Al considerar las frecuencias relativas en la tabla de contingencia se obtiene una tabla de frecuencias relativas denominada **distribución conjunta de frecuencias relativas**.

	1. White	2. Black	3. Asian	4. Other	Suma
1. < HS Grad	0.070	0.010	0.005	0.004	0.089
2. HS Grad	0.274	0.035	0.010	0.004	0.324
3. Some College	0.177	0.031	0.006	0.003	0.217
4. College Grad	0.192	0.013	0.022	0.001	0.228
5. Advanced Degree	0.113	0.008	0.020	0.001	0.142
Suma	0.827	0.098	0.063	0.012	1.000

Cuadro 3: Distribución conjunta de frecuencias relativas. Educación por raza.

Datos cualitativos-cualitativos

Representación tabular. Distribución conjunta de frecuencias relativas

De manera general

F	1	...	j	...	p	Suma
1	f_{11}	...	f_{1j}	...	f_{1p}	$f_{1.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	f_{i1}	...	$f_{ij} = \frac{k_{ij}}{n}$	...	f_{ip}	$f_{i.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	f_{n1}	...	f_{nj}	...	f_{np}	$f_{n.}$
Suma	$f_{,1}$	...	$f_{.j}$	...	$f_{.p}$	1

Cuadro 4: Distribución conjunta de frecuencias relativas. Forma general.

Datos cualitativos-cualitativos

Representación tabular

Al considerar las frecuencias relativas en la tabla de contingencia **respecto a los totales columna** se obtiene una tabla denominada **distribución condicionada columna de frecuencias relativas** ó **perfiles columna**.

	1. White	2. Black	3. Asian	4. Other
1. < HS Grad	0.085	0.106	0.079	0.297
2. HS Grad	0.331	0.358	0.163	0.351
3. Some College	0.215	0.314	0.095	0.216
4. College Grad	0.232	0.137	0.347	0.081
5. Advanced Degree	0.137	0.085	0.316	0.054
Suma	1.000	1.000	1.000	1.000

Cuadro 5: Distribución condicional columna de frecuencias relativas. Educación por raza.

Datos cualitativos-cualitativos

Representación tabular

Notas

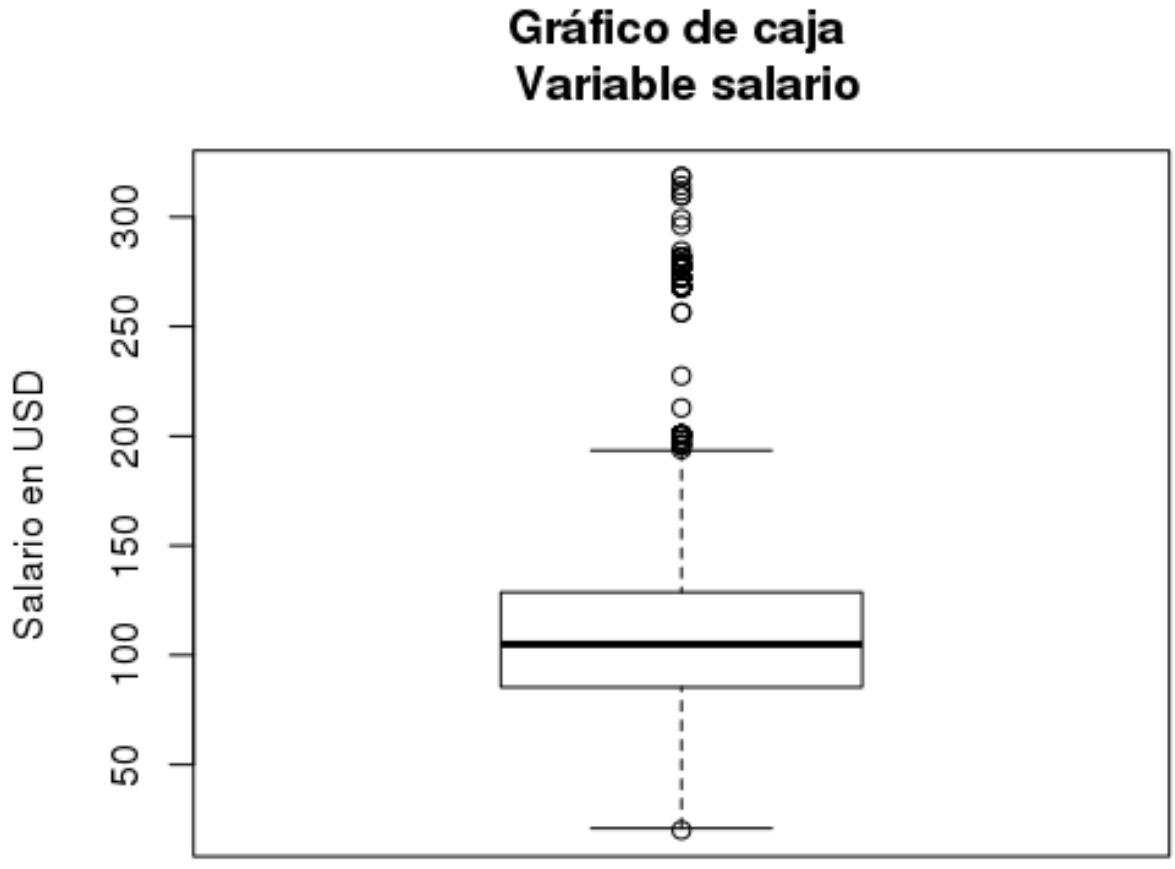
- El total de **K** es siempre igual a n.
- El total de **F** es siempre igual a 1.
- Las sumas por filas de **K** da como resultado los **marginales filas**. Similar para **F**.
- Las sumas por columnas de **K** da como resultado los **marginales columnas**. Similar para **F**.
- Así como perfiles columna, puedo encontrar los **perfiles fila** para una tabla de contingencia **K**.

Datos cualitativos-cuantitativos

Además de la variable *escolaridad* para la muestra de $n = 3000$ personas, se cuenta con la variable *salario*. Anteriormente observamos la distribución de esta variable. ¿cómo se relaciona con la variable *escolaridad*?

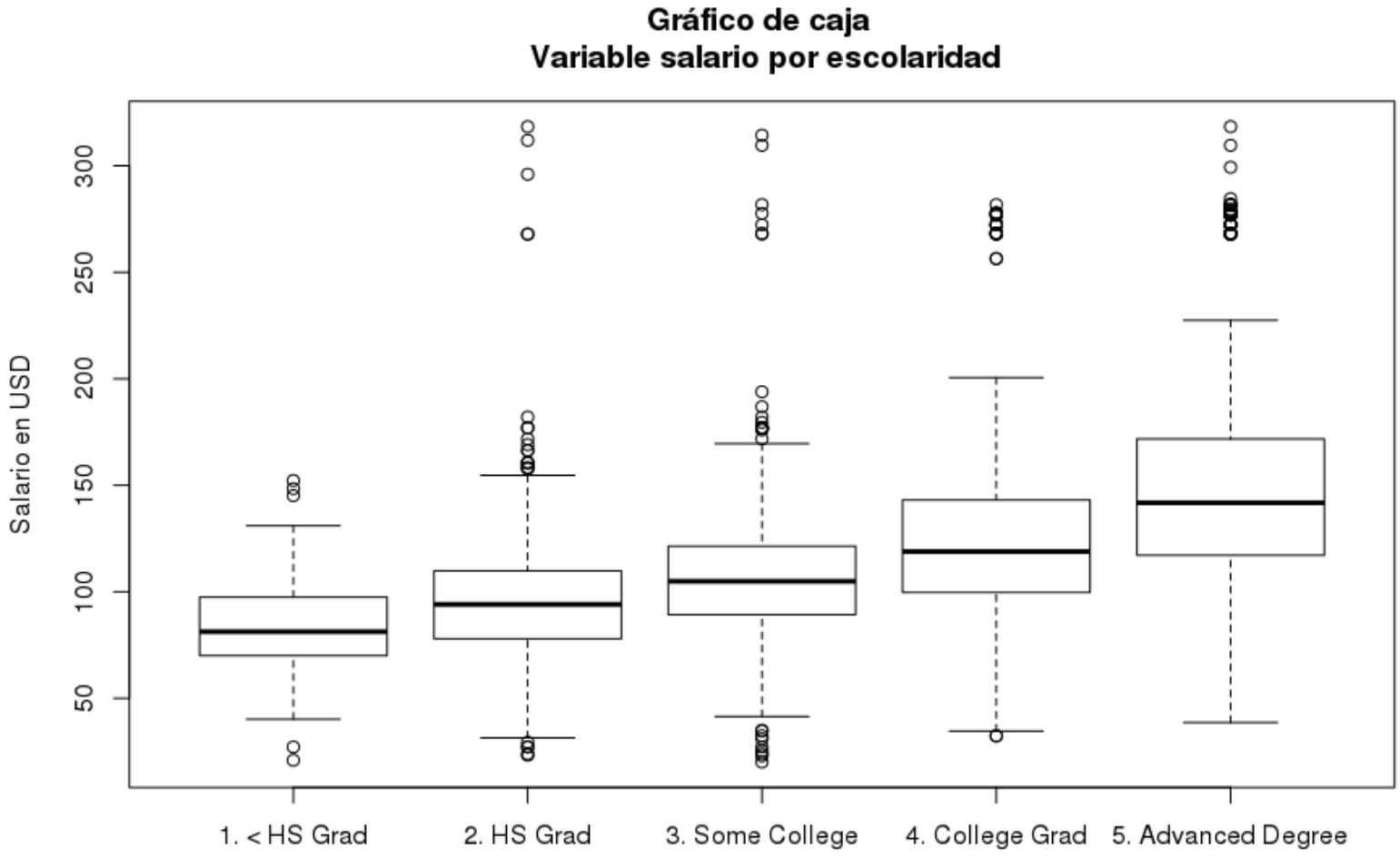
Datos cualitativos-cuantitativos

Figura 7: Gráfico de caja para la variable *salario*



Datos cualitativos-cuantitativos

Figura 8: Gráfico de caja para la variable *salario* en los niveles de la variable *escolaridad*



Datos cualitativos-cuantitativos

¿Cómo se relaciona el *salario* con la variable *escolaridad* para las diferentes *razas*?

- En realidad es un problema de tres variables.
- Una tabla de contingencia dónde las filas representen la escolaridad, las columnas las razas y sus entradas el salario promedio correspondiente a la escolaridad $\times$ raza.
- Esto último puede ser graficado mediante un gráfico de barras lado a lado.

Contenido

Presentación gráfica para datos cuantitativos-cuantitativos

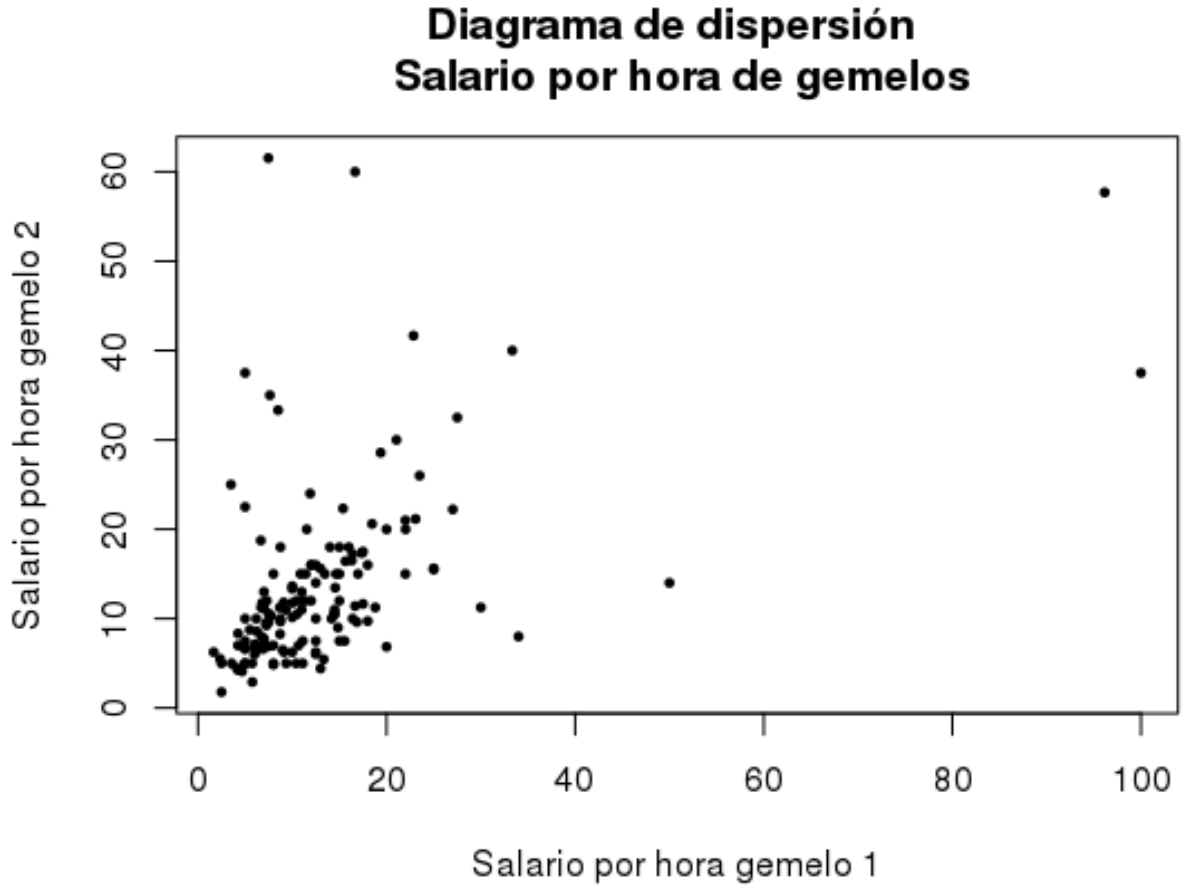
¿Cómo se presentan los datos?

Cuando las dos variables que hayan de presentarse en una gráfica son cuantitativas, una de ellas se grafica a lo largo del eje horizontal (denotado como x) y la otra a lo largo del eje vertical (denotado como y). Cada par de valores de datos se grafica como punto en esta gráfica de dos dimensiones, llamada **gráfica de dispersión**.

Presentación gráfica para datos cuantitativos-cuantitativos

¿Cómo se presentan los datos?

Figura 9: Gráfico de dispersión para las variables *salario por hora de gemelo 1*, *x* y *salario por hora de gemelo 2*, *y*



Presentación gráfica para datos cuantitativos-cuantitativos

¿Cómo se presentan los datos?

- ¿Qué tipo de modelo se muestra?
  Creciente, decreciente, constante, lineal recto, lineal curvo ó simplemente aleatorio.
- ¿Qué tan fuerte es el modelo?
  ¿Es el modelo observado es seguido por todos los puntos de forma exacta ó es débilmente evidenciable?
- ¿Hay observaciones poco comunes ó agrupaciones?
  ¿Hay una o varias observaciones lejanas al modelo observado? ¿existen agrupaciones de observaciones? ¿hay alguna explicación para estos valores?

Contenido

Medidas numéricas para datos cuantitativos-cuantitativos

¿Cómo se cuantifica la relación de los datos?

Una tasa constante de aumento o disminución es quizá el modelo más común que se encuentra en gráficas de dispersión bivariadas. Cuando éste es el caso, decimos que las dos variables exhiben una **relación lineal**.

Medidas numéricas para datos cuantitativos-cuantitativos

¿Cómo se cuantifica la relación de los datos?

Para una relación lineal surgen las siguientes preguntas:

1. ¿cómo cuantificar si **existe** dicha relación lineal?.
   Sabemos caracterizar numéricamente cada variable de manera univariada ($\bar{x}$, s_x y $\bar{y}$, s_y) pero no su relación bivariada.
2. ¿cómo diferenciar si la relación lineal es **positiva** ó **negativa**?
   Si el modelo es lineal ¿como cuantificar su dirección? ($\nearrow$ ó $\searrow$)
3. ¿cómo determinar la **fuerza** de la relación lineal?
   Si la relación es fuerte ó débil ¿cómo diferenciarla?

Medidas numéricas para datos cuantitativos-cuantitativos

¿Cómo se cuantifica la relación de los datos?

Para un conjunto de n mediciones de x: $x_1,\ldots,x_n$ y n mediciones de y: $y_1,\ldots,y_n$ se define la covarianza muestral s_{xy} como:

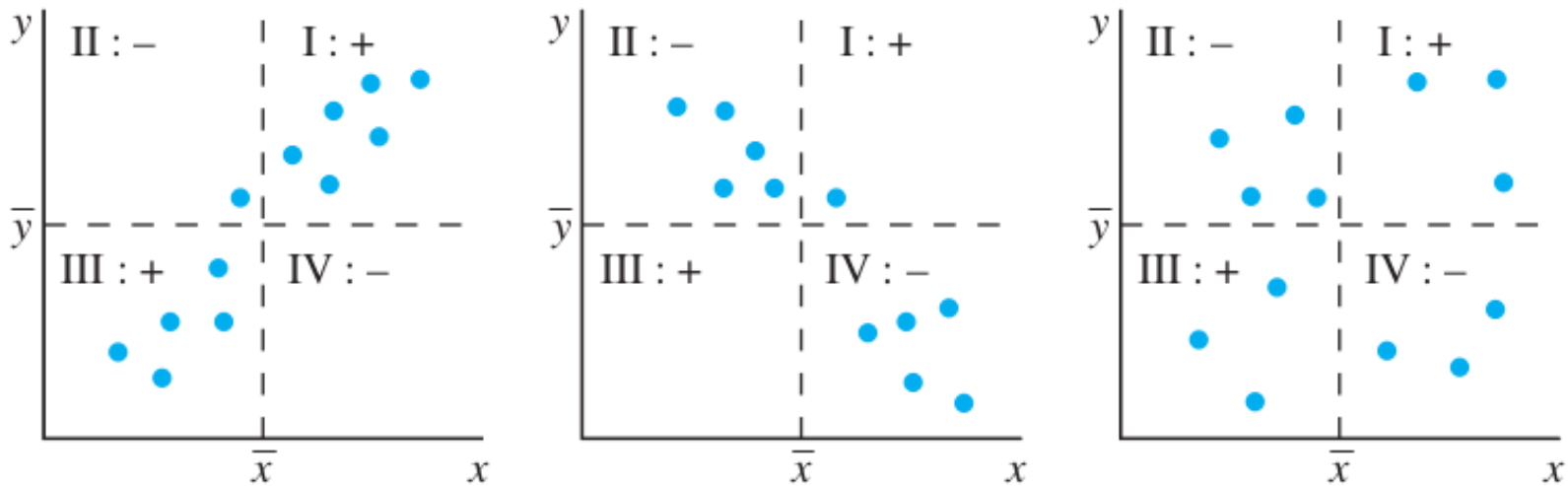
$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n}}{n-1}$$

Medidas numéricas para datos cuantitativos-cuantitativos

¿Cómo se cuantifica la relación de los datos?

Observemos que la covarianza s_{xy} nos permite responder las primeras 2 preguntas.

Figura 10: Tipos de relaciones lineales con signos del producto cruz $(x_i - \bar{x})(y_i - \bar{y})$. Tomado de Introducción a la probabilidad y Estadística.



Medidas numéricas para datos cuantitativos-cuantitativos

¿Cómo se cuantifica la relación de los datos?

Para un conjunto de n mediciones de x: $x_1,\ldots,x_n$ con desviación estándar muestral s_x y n mediciones de y: $y_1,\ldots,y_n$ con desviación estándar muestral s_x se define el coeficiente de correlación muestral r como:

$$r = \frac{s_{xy}}{s_x s_y}$$

Conclusiones

- Al describir de manera simultánea dos variables, es relevante caracterizar la asociación que estas presentan.
- De manera semejante al escenario univariado, se presenta la descripción bivariada desde un punto de vista gráfico y numérico. En particular, la covarianza y correlación para el caso cuanti-cuanti fue estudiado.

Ejercicio de aplicación

Quinto laboratorio de programación en R

Realice el quinto laboratorio de programación en R.