# Trabajo Eje 4 Estadística Descriptiva

Julian Camilo Gonzalez Quevedo
Paola Andrea Mejia Bedoya
Carolay Helena Meza Mesa
Ingrid johanna Segura Urbaneta

Noviembre 2020

Fundación Universitaria del Área Andina
Administración de Mercadeo
Estadística descriptiva

## ***Introducción***

Por medio de este trabajo, esperamos cimentar las bases de lo aprendido en clase sobre cómo abordar cuantitativamente el análisis de la relación lineal entre dos variables continuas. Este tipo de relaciones son básicas ya que están presentes en todo tipo de análisis que debemos abordar: por ejemplo, ¿en cuánto aumentan las ventas con cada millón de pesos gastado en marketing? Este es uno de los tipos de preguntas que podemos estudiar estadísticamente con las siguientes tres herramientas estadísticas: el coeficiente de correlación, el diagrama de dispersión y la regresión lineal.

Para la práctica de estas herramientas, utilizaremos una base de datos de 21 ciudades estadounidenses. Para cada ciudad, tenemos dos variables: ***(i)*** *el porcentaje de conductores menores de 21 años;* ***(ii)*** *el número de accidentes fatales por* $1,000$ *licencias.* A lo largo del trabajo, intentaremos estudiar la relación entre estas dos variables. En particular, intentaremos entender qué efecto tiene el porcentaje de conductores adolescentes sobre el índice de accidentes. Nuestra primera hipótesis será que los conductores adolescentes presentan un mayor riesgo de seguridad vial que los conductores más experimentados. De acuerdo a esta hipótesis, esperamos encontrar una correlación positiva y significativa entre las dos variables; es decir, esperamos que las ciudades con mayor porcentaje de conductores adolescentes también tengan un mayor número de accidentes de tránsito fatales.

Finalmente, estudiaremos la validez estadística de esta hipótesis a través de las tres herramientas anteriormente nombradas: el coeficiente de correlación, el diagrama de dispersión y la regresión lineal.

## ***Datos***

Nuestro primer paso será cargar los datos desde excel al lenguaje de programación estadístico R.

**## # A tibble: 21 x 2**
**## porcentajes_de_menores_de_21_anos accidentes_fatales_por_1000_licencias**
**## <dbl> <dbl>**

| PORCENTAJES DE MENORES DE 21 AÑOS | ACCIDENTES FATALES POR 1000 LICENCIAS |
|---|---|
| 13 | 2,962 |
| 12 | 0,708 |
| 8 | 0,885 |
| 12 | 1,652 |
| 11 | 2,091 |
| 17 | 2,627 |
| 18 | 3,83 |
| 8 | 0,368 |
| 13 | 1,142 |
| 8 | 0,645 |
| 9 | 1,028 |
| 16 | 2,801 |
| 12 | 1,405 |
| 9 | 1,433 |
| 10 | 0,039 |
| 9 | 0,338 |
| 11 | 1,849 |
| 12 | 2,246 |
| 14 | 2,855 |
| 14 | 2,352 |
| 11 | 1,294 |

Después de cargar los datos y limpiar los nombres de las variables, terminamos con los siguientes datos para 21 ciudades estadounidenses.

- El porcentaje de conductores que tienen menos de 21 años. Esta variable será nuestra variable X en el análisis: es decir, nuestra variable explicativa.
- El número de accidentes fatales por cada $1,000$ licencias en cada ciudad. Estas variables serán, nuestra variable Y en el análisis: es decir, nuestra variable a explicar

a partir de la variable explicativa (Porcentaje de conductores que tienen menos de 21 años).

## ***Coeficiente de Correlación***

A continuación debemos hallar los valores del denominador y numerador de la ecuación del coeficiente de correlación para esto debemos encontrar los datos que nos pide dicha ecuación.

$$r = \frac{n \cdot \sum x_i \cdot y_i - \sum x_i \cdot \sum y_i}{\sqrt{[n \cdot \sum x_i^2 - (\sum x_i)^2] \cdot [n \sum y_i^2 - (\sum y_i)^2]}}$$

$$-1 \leq r \leq 1$$

- Para esto tomando los valores de la tabla anterior realizamos los cálculos correspondientes, en este caso (X*Y), (X ^2) Y (Y^2),
- Debemos tener en cuenta que para realizar dicha ecuación tenemos que nuestra variable **n = 21**

| | X*Y | X^2 | Y^2 |
|---|---|---|---|
| | 38,506 | 169 | 8,773444 |
| | 8,496 | 144 | 0,501264 |
| | 7,08 | 64 | 0,783225 |
| | 19,824 | 144 | 2,729104 |
| | 23,001 | 121 | 4,372281 |
| | 44,659 | 289 | 6,901129 |
| | 68,94 | 324 | 14,6689 |
| | 2,944 | 64 | 0,135424 |
| | 14,846 | 169 | 1,304164 |
| | 5,16 | 64 | 0,416025 |
| | 9,252 | 81 | 1,056784 |
| | 44,816 | 256 | 7,845601 |
| | 16,86 | 144 | 1,974025 |
| | 12,897 | 81 | 2,053489 |
| | 0,39 | 100 | 0,001521 |
| | 3,042 | 81 | 0,114244 |
| | 20,339 | 121 | 3,418801 |
| | 26,952 | 144 | 5,044516 |
| | 39,97 | 196 | 8,151025 |
| | 32,928 | 196 | 5,531904 |
| | 14,234 | 121 | 1,674436 |
| TOTAL | 455,136 | 3073 | 77,451306 |

Con estos datos, deseamos entender si hay alguna relación estadística entre el porcentaje de conductores que tienen menos de 21 años y el número de accidentes fatales por cada $1,000$ licencias. Es decir, queremos ver si hay fundamento estadístico para la siguiente hipótesis: ¿Se presentan más accidentes de tránsito fatales en las ciudades donde hay mayor porcentaje de conductores adolescentes?

Un primer análisis para responder a esta pregunta es el coeficiente de correlación de Pearson. Esta variable estadística mide la co-relación lineal entre dos variables. Por ejemplo, sus valores están acotados entre -1 e 1. Luego, si hay una correlación de 1 entre una variable X y Y, implica que hay una correlación perfecta entre ellas: si aumenta X, aumenta Y igualmente. Por otro lado, si hay una correlación de -1, tenemos la situación opuesta: si aumenta X, disminuye Y.

Dentro de nuestro ejemplo, esperamos una correlación positiva entre el porcentaje de conductores que tienen menos de 21 años y el número de accidentes fatales por cada $1,000$ licencias. Es decir, esperamos que entre mayor sea el porcentaje de conductores adolescentes, mayor sea el número de accidentes fatales.

A continuación, calculamos el coeficiente de correlación de Pearson para las siguientes variables:

- El porcentaje de conductores que tienen menos de 21 años en cada ciudad.
- El número de accidentes fatales por cada $1,000$ licencias en cada ciudad.

```
cor(datos$porcentajes_de_menores_de_21_anos,datos$accidentes_fatales_por_1000_licencias)
```

```
## [1] 0.829189
```

Nuestros resultados arrojan ***un coeficiente de correlación de 0.829.*** Es decir, hay una correlación positiva y de gran magnitud entre el porcentaje de conductores que tienen menos de 21 años y el número de accidentes fatales por cada $1,000$ licencias. Por lo tanto, concluimos que hay evidencia estadística que refuerza nuestra ***hipótesis: Efectivamente, se presentan más accidentes de tránsito fatales en las ciudades donde hay mayor porcentaje de conductores adolescentes.***

Concluimos, por lo tanto, que ***los conductores jóvenes representan un riesgo de accidente mayor que sus contrapartes más experimentadas.*** De igual manera, estos riesgos se reflejan en las estadísticas por ciudad al comparar sus índices de accidentes y sus porcentajes de conductores adolescentes.

## ***Diagrama de dispersión***

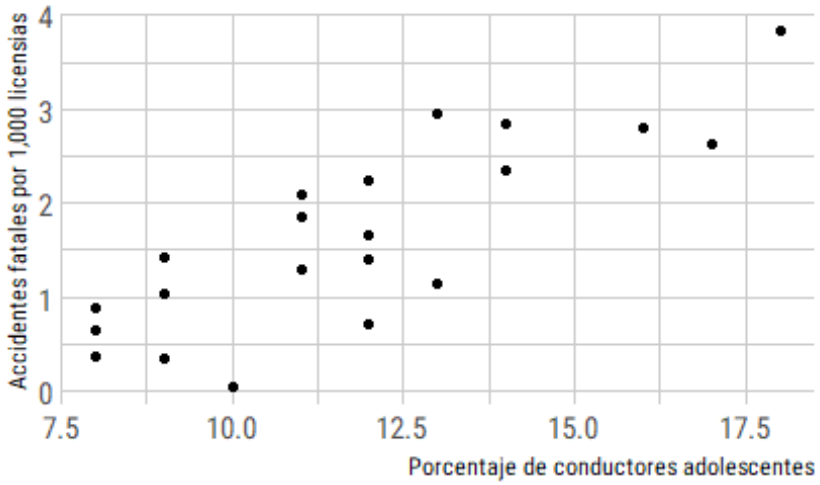
Por otro lado, podemos seguir evaluando nuestra hipótesis, pero esta vez con un análisis gráfico: al graficar en un plano cartesiano el porcentaje de conductores que tienen menos de 21 años y el número de accidentes fatales por cada $1,000$ licencias, podremos confirmar si el número de accidentes aumentan en las ciudades que tienen un mayor porcentaje de conductores adolescentes.

A continuación, graficamos el diagrama de dispersión. En el eje y, estará el número de accidentes; en el eje x, el porcentaje de conductores adolescentes.

```r
datos %>%
  ggplot(aes(porcentajes_de_menores_de_21_anos, accidentes_fatales_por_1000_licencias)) +
  geom_point() +
  labs(x = "Porcentaje de conductores adolescentes",
       y = "Accidentes fatales por 1,000 licencias",
      title = "Relación lineal positiva entre las dos variables",
      subtitle = "A mayor porcentaje de conductores adolescentes, mayor número de accidentes")
```

***RELACIÓN LINEAL POSITIVA ENTRE LAS DOS VARIABLES***

***A mayor porcentaje de conductores adolescentes, mayor número de accidentes***




Tal y como lo esperábamos, podemos confirmar una correlación gráfica positiva entre las dos variables de interés. Es decir: las ciudades con mayor porcentaje de conductores adolescentes presentan un número mayor de accidentes. Esta gráfica valida nuestro análisis de correlación realizado en el punto anterior.

***Ecuación de la línea recta***

En primer lugar hallaremos los valores para nuestra pendiente, buscarnos el ***numerador 1024,01*** y ***denominador 3524,00***, al realizar la división. nuestra pendiente **= *0,29.***

En segundo lugar para hallar los valores de nuestrintercepto, tenemos que **B= Y -AX,** al remplazar los datos correspondientes, tenemos que ***Y = 1,65,*** recordemos que este es el promedio de eje Y, **A= 0,29 * X=11,76** este es el promedio de X**,** al resolver esta ecuación nos arroja que el valor del intercepto **B es - 1,77**

| | | |
|---|---|---|
| Numerador | 1024,01 | ECUACIONES |
| Denominador | 3524,00 | |
| Pendiente (A) | 0,29 | A= (N∑X.Y)-(∑X).(∑Y) / (N∑X^2) - (∑X)2 |
| Intercepto (B) | -1,773 | B:Y-AX |
| Ecuaciòn de la linea recta | y= Ax + B | |
| | y=0.29x-1.77 | |

Con la anterior tabla mostramos que nuestra ecuación de la línea recta es

**Y= Ax + B** reemplazando valores esto queda ***Y= 0.29x - 1.77***

## ***Regresión***

Finalmente, para analizar nuestra hipótesis, tenemos un tercer método fuertemente ligado a los anteriores: una regresión lineal. Con este método, esperamos responder a la siguiente pregunta: ¿por cada cambio de una unidad en la variable explicativa, en este caso el porcentaje de conductores adolescentes, cuánto esperamos que cambie el índice de accidentes por $1,000$ personas?

En línea con nuestros resultados anteriores, esperamos que haya una relación positiva: un punto porcentual más en el porcentaje de conductores adolescentes debería estar asociado con un mayor índice de accidentes. En términos matemáticos, estimaremos la siguiente ecuación lineal:

$$accidentes = \beta_0 + \beta_1(\%conductores\ adolescentes)$$

```
lm(accidentes_fatales_por_1000_licencias ~porcentajes_de_menores_de_21_anos , data = datos)

##
## Call:
##  lm(formula = accidentes_fatales_por_1000_licencias ~ porcentajes_de_menores_de_21_anos,
##     data = datos)
##
## Coefficients:
##                   (Intercept)  porcentajes_de_menores_de_21_anos
##                       -1.7725                             0.2906
```

Nuestros resultados son los siguientes: para el intercepto, estimamos un coeficiente de -1,77 y un coeficiente para la variable explicativa (porcentaje de conductores que tienen menos de 21 años en cada ciudad) de 0.29. Es decir, esperamos que un aumento de 10 puntos porcentuales en el porcentaje de conductores menores a 21 esté asociado con un aumento de 2.9 accidentes fatales por cada $1,000$ licencias en cada ciudad.

Por lo tanto, este último hallazgo corrobora lo que habíamos evidenciado anteriormente: hay una correlación positiva y significativa entre el porcentaje de conductores adolescentes y el número de accidentes por $1,000$ personas. Con este último hallazgo gracias a nuestra regresión lineal, podemos estimar un número preciso para esta relación: un aumento de 10 puntos

porcentuales en el porcentaje de conductores menores a 21 esté asociado con un aumento de 2.9 accidentes fatales por cada 1, 000 licencias en cada ciudad.

Profesor Oscar, adjuntamos el ejercicio realizado en excel https://drive.google.com/file/d/12Uzo9rwXHMp73FkWAhYjGQTENx7xdlq8/view?usp=sharing .

## ***Conclusión***

Para concluir, por medio de este taller, hemos afinado nuestra habilidad estadística para el estudio de relaciones lineales entre variables continuas. A través de el coeficiente de correlación lineal, el diagrama de dispersión y la regresión lineal, hemos estudiado la relación entre el porcentaje de conductores que tienen menos de 21 años y el número de accidentes fatales por cada $1,000$ licencias. Al comienzo de este trabajo, teníamos la siguiente hipótesis: los conductores adolescentes presentan un mayor riesgo de seguridad vial que los conductores más experimentados.

Utilizando tanto el coeficiente de correlación lineal, el diagrama de dispersión y la regresión lineal, hemos concluido que nuestra hipótesis es correcta. Efectivamente, el porcentaje de conductores adolescentes está positivamente correlacionado con el número de accidentes en la ciudad. En particular, al estimar la regresión lineal, concluimos que un aumento en un punto porcentual del porcentaje de conductores adolescentes está asociado con 0.29 más accidentes por 1,000 licencias. Es decir, las ciudades donde manejan un mayor porcentaje de adolescentes también presentan mayores índices de accidentes.

Con estos hallazgos, realizamos las siguientes dos recomendaciones que podrían disminuir el número de accidentes de tráfico al desincentivar que los adolescentes manejan.

1. La primera recomendación es hacer más caros los trámites para que los adolescentes consignan su licencia de conducir. Por lo tanto, menos adolescentes conseguirán su licencia, menos adolescentes manejarán y, por ende, habrá un menor número de accidentes.

2. La segunda recomendación es crear una campaña de concientización en los jóvenes. A través de diversas campañas educativas, se podría concientizar a los jóvenes sobre los

peligros de conducir sin responsabilidad. Por lo tanto, los jóvenes ya no serán una mayor fuente de riesgo vial y, por ende, habrá un menor número de accidentes.

Para concluir, en este taller hemos evidenciado cómo el análisis estadístico puede ayudarnos a entender cuantitativamente diversos fenómenos. En este caso, hemos podido ver como el coeficiente de correlación, el diagrama de dispersión y la regresión lineal nos han ayudado a entender mejor el fenómeno de accidentalidad vial en las poblaciones adolescentes en Estados Unidos. De esta manera, hemos podido formular dos recomendaciones que disminuyan el índice de accidentalidad.

## ***LISTA DE REFERENCIAS***

Castillo, P. (2020). ESTADÍSTICA  Eje 4. Fundación Universitaria Área Andina.