Introducción a la Ciencia de Datos

Guía de trabajos prácticos N°3

Distribuciones

En esta guía vamos a trabajar con el dataset de costos de seguros que usaron para la primera entrega para introducir el concepto de **distribución** y discutir algunas propiedades. En el camino, vamos a presentar varias herramientas para graficar: `geom_dotplot`, `geom_histogram`, `geom_density`, `geom_freqpoly`, y otros.

Para este trabajo, van a tener como guía un archivo .R que está en el campus (distribuciones.R)

Parte 1. Preparación de los datos

1. Como siempre, abran RStudio y comiencen un nuevo script (**Ctrl+Shift+N** o en el menú de arriba a la izquierda) sobre el cual van a trabajar.
2. Carguen la biblioteca 'tidyverse' y lean el dataset de `insurance.csv`.
3. Filtren la tabla resultante para quedarse solo con los clientes hombres fumadores (`sex=='male'`, `smoker=='yes'`) ¿Cuántas filas tiene el dataset resultante?

Parte 2. *Dot plot.*

4. Como recordatorio hagan un gráfico de puntos de los gastos médicos (charges) en función de la edad de lxs aseguradxs. Esto debería salir fácil a esta altura, pero si necesitan ayuda pueden ver el código del campus como guía. ¿Qué pueden decir del costo del seguro de los fumadores a partir de este gráfico?

5. **Usando el código** del campus como guía, vamos a estudiar la *distribución* de los cargos médicos de la muestra filtrada. Para esto vamos a construir un gráfico similar al que vimos en la parte expositiva de la clase. Este gráfico tiene dos elementos:

    a. Un diagrama de dispersión (scatter plot; usando `geom_point` o `geom_jitter`) de los costos, según el eje x (usar jitter para que se vean los puntos).
    b. Un diagrama de puntos sobre el eje horizontal.

6. Modificar el ancho de las "clases" (*bins*) en los que se divide el dataset con el argumento binwidth (ancho del bin). Aumentar y disminuir el ancho de banda. ¿Cómo cambia el gráfico según el tamaño del bin? ¿Cambia el número de puntos que hay en cada caso?

> **Nota**
>
> Para ver bien el gráfico, en algunos casos tal vez tengan que cambiar el argumento `dotsize` de `geom_dotplot`, que indica el tamaño de los puntos en función del ancho del bin.

Parte 3. Histogramas.

7. Repitan la gráfica de arriba, sin filtrar el dataset original. ¿Qué pasa? ¿Sirve el gráfico resultante? ¿cuál es una limitación del *dot plot*? En otras palabras, si tuvieran que decirle a alguien cuándo **no** usar este tipo de gráfico, ¿qué le dirían?

> **Importante**
>
> Un gráfico extremadamente popular, y que cumple una función similar que los *dot plots*, pero con la gran ventaja de que se pueden usar para datasets de tamaños arbitrarios son los famosos **histogramas**. Los histogramas también dividen al conjunto de datos en clases (bines), pero en lugar de graficar cada punto del conjunto, se hace una barra por clase, la altura de esta barra está relacionada con la cantidad de puntos que caen en esa clase. En ggplot, podemos construir fácilmente un histograma con `geom_histogram`.

8. Modifiquen el código (con el dataset filtrado) para usar un histograma en lugar de un dot plot. Fíjense que en este caso, las etiquetas eje y tienen información. Modifiquen el código:

    a. Volver a hacer visibles las etiquetas del eje y.

b. Que los gráficos no se superpongan.

> **Nota**
>
> Igual que en el *dot plot*, el ancho de los bins de los histogramas también puede modificarse con el argumento `binwidth`. En este caso, también existe el argumento bins, que permite indicar el número de clases, en lugar del ancho.

9. Cambien el ancho / número de bines y vean qué pasa con la gráfica. Presten atención a los valores del eje y. ¿Cómo cambian los valores del eje y cuando se aumenta el número de bines (se disminuye el ancho)? ¿Y cuando el número de bines disminuye?

10. Realicen un histograma de todos los cargos médicos (del dataset sin filtrar) utilizando un número de bines adecuado. Respondan las siguientes preguntas:

    a. ¿Qué variable se representa en el eje X?
    b. ¿Qué representa el eje Y en este gráfico?
    c. ¿Cuál es el rango de valores de los cargos del seguro de salud?
    d. ¿Cómo describirían la distribución de los cargos del seguro de salud? (Ejemplo: simétrica, sesgada a la derecha, sesgada a la izquierda)
    e. ¿Qué podemos inferir sobre el costo del seguro de salud a partir de este histograma?
    f. Si se hiciera un histograma solo con fumadores, ¿qué diferencias esperarías encontrar respecto a este gráfico?

11. Basándose en el gráfico del punto anterior, respondan Verdadero o Faslo:

    a. La mayoría de las personas pagan costos de seguro de salud superiores a 50,000.
    b. Un número menor de bins haría que el histograma mostrara menos detalle sobre la distribución de los datos.
    c. Un menor tamaño de bin haría que el histograma mostrara menos detalle sobre la distribución de los datos.
    d. Si hiciéramos el histograma con datos de no fumadores, necesariamente tendría la misma forma.

Parte 4. Densidades.

Investiguemos un poco los subproductos de hacer un histograma. Obviamente, la altura de las barras se corresponden con la cantidad de *datapoints* en ese bin. Como ya

hablamos, para cada `geom_` tenemos una función `stat_` correspondiente, que hace las cuentas. En el caso de `geom_histogram`, la función es `stat_bin`.

Podemos recuperar ese valor de cuentas por bin con el siguiente código (un poco rebuscado):

```
g <- ggplot(data=df) + stat_bin(aes(x=charges))
a <- ggplot_build(g)
datos <- a[1]$data[[1]]
```

La tabla datos contiene toda la información que se usa para generar el histograma, incluyendo una columna count que es lo que estamos buscando:

```
datos$count
```

Podemos usar esta información para agregar etiquetas en el histograma que hicimos arriba. Para esto, usamos `stat` con el `geom` "text", y accedemos a las columnas de datos rodeando a los nombres con dos puntos. Ver el código de guía.

12. Podemos usar otras características para mapear el dataset. Volvamos al dataset original, filtremos solo varones (pero dejemos fumadores y no fumadores). Usen la variable `smoker` para mapear al color de relleno de los histogramas (`fill`) y usen `position='identity'` para evitar que ponga una barra arriba de la otra, como veníamos haciendo. Vean el código para ayudarse, si necesitan.

13. **Respondan**: ¿sirve este gráfico para comparar cómo difieren los gastos de seguro entre varones fumadores y no fumadores? ¿Cuál es el tamaño de cada uno de los grupos?

> **!** Importante
>
> Si queremos comparar la distribución de los valores de una variable (en este caso, los costos del seguro) para dos grupos que no tienen la misma cantidad de muestras, se vuelve necesario normalizar de alguna manera la altura de las barras. Lo primero que se nos puede ocurrir, es dividir por la altura de la barra más alta para cada histograma, o dividir por la cantidad de muestras en cada grupo.
> Sin embargo, una forma más inteligente de hacerlo y que tiene más sentido probabilístico, es dividir por la cantidad de muestras y el ancho de los bines. De esta manera, convertimos al eje y en una ***densidad*** de puntos. Es decir, obtenemos valores tal que la *integral* del histograma de principio a fin da como resultado 1, independientemente de la cantidad de muestras de cada clase.

Esta densidad es calculada por el `stat_bin`, como pueden ver explorando el objeto `datos` que definimos más arriba (en la columna "density").

14. Modifiquen el código anterior, poniendo como argumento del `geom_histogram` un mapeo a la variable `..density..`, que sale del stat correspondiente (recordar la parte del código en el que hicimos el etiquetado de las barras del histograma más arriba). ¿Cómo difiere este histograma del realizado en el punto 11?

i Nota

Un histograma normalizado de esta manera es una forma de estimar la *función de densidad de probabilidad* (PDF) de una variable continua a partir de un conjunto de datos. Las PDF contienen la máxima información que uno puede obtener de una variable. A partir de ellas uno puede calcular estadísticos de resumen, obtener datos simulados, etc.

Otra forma de estimar la PDF es usar una **estimación de densidad con kernel**. Sin necesidad de entrar en los detalles técnicos de cómo se calcula (algo vimos en la parte expositiva), podemos utilizar este tipo de estimación a partir del `geom_density`. Esta función calcula y grafica la estimación de densidad a partir de un conjunto de datos. Su uso es muy similar al de los otros geom, tipo `geom_histogram`.

15. Agreguen al gráfico del punto anterior una estimación de densidad de los mismos datos, sumando un `geom_density`. Piensen si necesitamos hacer algún mapeo extra o no. Analicen el resultado. ¿Entienden intuitivamente lo que está haciendo la estimación de densidad por kernel? Comparen su intuición con alguien que esté trabajando en otra computadora.

16. La estimación de densidad por kernel tiene muchos parámetros, pero tal vez el más relevante es el ancho de banda (*bandwidth*; no confundir con `binwidth`), que representa el área de influencia de un punto en la estimación de la densidad. Cambien este parámetro y vean cómo se modifica el gráfico de la estimación de densidad. Prueben varias opciones ancho de banda diferentes y elijan la que les parezca que mejor describe los datos. Un poco de introspección: ¿qué los lleva a elegir un ancho de banda sobre otro?

17. Realicen un gráfico de densidad que sirva para comparar las distribuciones de cargos médicos de las mujeres fumadoras y las no fumadoras. Respondan las siguientes preguntas:

    a. ¿Qué representa el eje X en este gráfico?

b. ¿Cuál es la diferencia entre las distribuciones de cargos médicos para fumadoras y no fumadoras?
c. ¿Qué se puede decir sobre la *variabilidad* de los costos del seguro para fumadoras en comparación con no fumadoras?
d. ¿Por qué la densidad de las no fumadoras es mayor en los valores más bajos de costos? ¿Qué implica esto?
e. ¿Cómo cambiaría la interpretación del gráfico si en lugar de una función de densidad se usara un histograma?

18. Basándose en el gráfico del punto anterior, respondan Verdadero o Faslo:

    a. El eje Y representa la cantidad total de mujeres en cada categoría.
    b. La densidad de las no fumadoras es mayor en valores de costos bajos.
    c. La mayoría de las fumadoras paga costos de seguro menores a 15,000.
    d. Si eliminàramos a las fumadoras del gráfico, la distribución de las no fumadoras cambiaría.
    e. La densidad para las fumadoras es mayor en valores altos de "charges" en comparación con las no fumadoras.
    f. La densidad acumulada de ambas distribuciones sumará exactamente 1.

Parte 5. Comparaciones múltiples.

Hasta ahora, hicimos histograma y gráficos de densidad comparando dos categorías. Sin embargo, cuando hay más de dos o tres, interpretar ese tipo de gráfico se vuelve complicado.

19. Prueben comparar las distribuciones de costos de seguro de cada una de las cuatro regiones (variable region) en un solo gráfico de histogramas, mapeando el color (o el `fill`, usando algún valor de alpha por debajo de 0.5).

    > **Nota**
    >
    > Recuerden que por defecto, los histogramas apilan, y nosotros queremos verlos uno arriba del otro para comparar las distribuciones, por lo que tienen que fijar `position='identity'`.

Una alternativa es usar `geom_freqpoly`, que en lugar de graficar barras para las cuentas en los bines, las une con líneas. Implementen el gráfico para comparar las regiones usando los polígonos de frecuencias. Noten que en este caso, el valor de la posición por defecto es `'identity'`. Esto ya debería ser indicador de para qué sirve cada uno de los gráficos.

Al igual que `geom_histogram`, este tipo de gráfica también tiene argumentos `binwidth`, `bin`, y también calcula variable `density`, que podemos usar para mapear a la coordenada vertical si queremos un gráfico expresado en densidades (importante si los conjuntos a comparar son muy disímiles en tamaño).

Parte 6 (extra). Más dimensiones

20. En algunos casos, vamos a querer ver los gráficos de densidad en función de varias variables categóricas. Una opción es usar la librería ggridges, que es un complemento de ggplot (seguramente tengan que instalarlo). Vean el código y jueguen para conseguir otros gráficos. También pueden revisar los numerosos ejemplos en la página del paquete.

21. En otros casos, vamos a querer ver la distribución de los valores de dos variables continuas a la vez (hasta ahora, solo estuvimos viendo la distribución de charges). Esto se conoce como la distribución *conjunta* de ambas variables. Existen varios tipos de gráficos que permiten ver esto. Exploren al menos los siguientes:

- `geom_bin_2d`
- `geom_hexbin`
- `geom_density_2d`

Bibliografía obligatoria

Libro de Wickham

Variación (histogramas): https://es.r4ds.hadley.nz/07-eda.html#variaci%C3%B3n

Covariación (polígonos de frecuencia): https://es.r4ds.hadley.nz/07-eda.html#covariaci%C3%B3n

R Graph Gallery

Histogramas: https://r-graph-gallery.com/histogram.html

Gráficas de densidad: https://r-graph-gallery.com/density-plot.html

Violines: https://r-graph-gallery.com/violin.html

Bibliografía para profundizar

GGridges: https://r-graph-gallery.com/ridgeline-plot.html