Introducción a la Ciencia de Datos - 2C 2022

Guía de Trabajos Prácticos Nº 4

Distribuciones

En esta guía vamos a trabajar con el dataset de costos de seguros que usaron para la primera entrega.

La clase pasada vimos cómo usar **métricas de resumen** para poder dar respuestas a ciertas preguntas menos directas usando datos. Sin embargo, también vimos que en ese resumen se pierde información. Hoy vamos a ver una posibilidad de dar una descripción más completa de los datos, pero que necesariamente será menos compacta. Vamos a 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*, *geom_violin*, 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. Limitaciones de las métricas de resumen y *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 reproducir el gráfico que vimos en la parte expositiva de la clase. Este gráfico tiene tres 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 boxplot (horizontal) por arriba de los puntos.

    c. Un diagrama de puntos sobre el eje horizontal.

¿Qué limitación encuentran en el diagrama de caja en este caso? ¿Qué característica clave de los datos **no** se captura con un *boxplot*? Responder esta última pregunta en el Formulario de la clase de hoy.

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. **Responder**: ¿Cómo cambia el gráfico según el tamaño del bin? ¿Cambia el número de puntos que hay en cada caso? Responder en el Formulario.

   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? **Responder**: ¿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? Responder esto último en el Formulario.

Una gráfica 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 con 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, cuya altura 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.

9. Igual que el *dot plot*, los bins de los histogramas también podemos modificarlos con el argumento `binwidth`. En este caso, también existe el argumento `bins`, que permite indicar el número de clases, en lugar del ancho. 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? Respondan estas últimas preguntas en el Formulario.

Parte 4. Densidades.

10. Investiguemos un poco los resultados subproductos de hacer un histograma. Obviamente, la altura de las barras se corresponden con la cantidad de *datapoints* del conjunto de datos 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 eso, 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.

11. 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.

12. **Respondan**: ¿sirve este gráfico para comparar cómo difieren los gastos de seguro entre mujeres y varones fumadores? ¿Cuál es el tamaño de cada uno de los grupos?

    Si queremos comparar la distribución de una propiedad (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").

13. 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?

Un histograma normalizado de esta manera es una forma de estimar la *función de densidad de probabilidad* (PDF) de una variable del dataset 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. Aclaramos que no es un tema de esta materia, pero es importante saber que se trata de algo muy valioso y con mucho interés.

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, 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`.

14. Agreguen al gráfico del punto trece 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.

15. La estimación de densidad por kernel tiene muchos parámetros, pero tal vez el más relevante es el ancho de banda (*bandwidth*), que representa el área de influencia de un punto en la estimación de la densidad. Cambien este parámetro y ver cómo se modifica el gráfico de la estimación de densidad. Prueben las opciones en el Formulario 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?

Nota sobre modelos?? Es uno de los primeros modelos que hacemos, no paramétrico. Sirve para predecir valores fuera del rango de los datos, por ejemplo.

Parte 5. Comparaciones múltiples.

16. En el punto 11 hicimos un histograma de dos categorías. Sin embargo, cuando hay más de dos o tres, interpretar ese tipo de gráfico. 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). 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).

17. Una alternativa interesante para comparar distribuciones de varios grupos, es el primo del gráfico de cajas: el gráfico de violines (`geom_violin`). La diferencia esencial con el gráfico de caja es que el de violines realiza estimaciones de densidad para cada grupo, como si fuera un `geom_density` para cada grupo. El gráfico de densidad aparece duplicado, de forma que crea algo parecido a un violín. Prueben crear un gráfico de violines con los costos del seguro según la región. Vean el código para ayudarse.

18. Una cosa que uno extraña inmediatamente, es la presencia de marcas que señalen los cuantiles más importantes. Pero `geom_violin` tiene un argumento `draw_quantiles`, que puede usarse para esto. Prueben con `draw_quantiles = c(0.25, 0.5, 0.75)`. Por supuesto, también pueden mapear otra variable a los colores de las líneas o de relleno de los violines. Pruébenlo.

Parte 6 (extra). Más dimensiones

19. En algunos casos, vamos a querer ver los gráficos de densidad en función de varias variables categóricas. Seguramente exploraron eso en el punto 18, pero a veces los gráficos pueden volverse un poco cargados, si seguimos agregando violines a la misma altura. 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.

20. 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`

Para trabajar en casa

Si aún no lo hiciste, contestá el Cuestionario de la Clase 4 antes de la clase virtual de esta semana.

Entrega 4. Análisis de las edades de los pasajeros del Titanic

Imaginen que son reclutados por una ONG que analiza datos históricos para hacer un reporte sobre las edades de los pasajeros del Titanic.

Pueden explorar la distribución según las clases, las tasas de supervivencia, etc. Dejen volar su imaginación. Pueden usar cualquier combinación de todas las herramientas que vimos hasta acá para hacer un reporte, **que entre en una carilla A4**. La presentación debe tener:

a. Un párrafo inicial donde se defina el objetivo, se cuente lo que se va a hacer y, si se animan, alguna hipótesis de lo que esperan encontrar.

b. Uno o dos gráficos, con un párrafo que acompañe a cada uno, y que incluya qué quieren analizar con ese gráfico, qué es lo que se muestra el gráfico y qué conclusiones que surgen del mismo.

c. Un párrafo en el que se concluya, relatando lo que se encontró y si se cumplió el objetivo planteado.

d. Si quisieran, también pueden incluir una tabla con métricas de resumen en algún lugar del documento.

Entreguen el informe en formato .pdf. Recuerden: debe tener una **extensión máxima de una carilla.**.