Clase 18-04-18: Análisis estadístico R

Ya conocemos las definiciones de indicador, de variable y sus tipologías. Hoy profundizaremos en conceptos sobre análisis estadístico.

- **Variable nominal**: En esta variable los valores no pueden ser sometidos a un criterio de orden, como por ejemplo los colores o el lugar de nacimiento.
- **Variable ordinal**: La variable puede tomar distintos valores ordenados siguiendo una escala establecida, aunque no es necesario que el intervalo entre mediciones sea uniforme.

Con este tipo de variables no es posible calcular medidas de tendencia central y medidas de dispersión. Las variables de intervalo y de razón permiten calcular estas medidas.

Dentro de la estadística existen medidas de tendencia central y de dispersión.

- **Medidas de tendencia central**: media, mediana y moda.
- **Media:** Es el promedio de los valores de una variable en una muestra. Resulta de la sumatoria de todos los valores dividida por el total de casos.
- **Mediana**: Es el valor que se encuentra en el centro de la distribución de una variable en una muestra.
- **Moda**: Es el valor más frecuente entre los valores de una variable en una muestra.

¿Por qué ocuparía la mediana en vez del promedio?

En una distribución normal la mediana coincide con el promedio. Sin embargo, estas distribuciones normales no son tan típicas. En algunos casos, el promedio puede dejar de ser representativo, entonces, en esos casos la mediana funciona mejor.

La moda no se utiliza tanto, pero en algunos casos permite ver dentro de los valores de la variable cuál es el que más se repite; y coincidentemente en una distribución normal también debería estar cerca de la mediana y el promedio.

Muchas de las reglas estadísticas tienen varios supuestos de entrada, que muchas veces no coinciden con los temas urbanos:

- Que las distribuciones son normales.
- Que las variables son independientes entre sí. En el espacio es muy difícil encontrar esto.

Nosotros estadísticamente podemos establecer comprobaciones de los fenómenos e incluso hablar de correlación, pero no causalidad.

Las medidas de tendencia central son un resumen de la información, pero no necesariamente dan cuenta de todos los aspectos de la distribución de una variable. Hay situaciones que no se pueden entender solo con el promedio, ya que tienen una distribución estándar muy alta. Ejemplo:

CASO 1:

Trabajador 1	$ 200.000
Trabajador 2	$ 230.000
Trabajador 3	$ 250.000
Trabajador 4	$ 320.000
Trabajador 5	$ 7.500.000
PROMEDIO	**$ 1.700.000**

CASO 2:

Trabajador 1	$ 1.600.000
Trabajador 2	$ 1.600.000
Trabajador 3	$ 1.700.000
Trabajador 4	$ 1.800.000
Trabajador 5	$ 1.800.000
PROMEDIO	**$ 1.700.000**

Las **medidas de dispersión** permiten expresar -a través de un solo indicador- de qué forma los valores están distribuidos en torno a un valor central:

- **Rango**: Es la medida de dispersión más sencilla, que mide la diferencia entre el valor más alto y el valor más bajo de la distribución. Me permite saber cuál es el dominio de la variable.

Ejemplo: Caso 1 → R= 1.500.000. Caso 2 → R= 100.000. La medida nos permite notar que hay algo raro en la medición.

- **Desviación estándar**: Es la medida de dispersión más utilizada. Resume las diferencias entre cada valor y el promedio. Mientras mayor sea la desviación, más heterogéneos serán los datos.

La varianza es lo mismo, pero sin el coeficiente al cuadrado.

$$s = \sqrt{\frac{\sum_{i=1}^{N}(Xi - \bar{X})^2}{N-1}}$$

- **Coeficiente de variación**: Permite comparar la desviación estándar de variables con diferentes unidades de medida.

Análisis estadístico bivariado

- **Variable dependiente**: Es la variable que se pretende explicar o caracterizar en la investigación.
- **Variable independiente**: Es la variable que conceptualmente se define como una posible "causa" o explicación de fenómeno a estudiar

Nivel de medición	Tipo de análisis	Prueba de asociación
Nominal - Nominal	Tabla cruzada	**Chi Cuadrado**; Lambda
Nominal - Ordinal	Tabla cruzada	Chi Cuadrado; Lambda
Ordinal - Ordinal	Tabla cruzada	Chi Cuadrado; Lambda; Gamma
Nominal - Numérica	Comparación de medias	T-Student; **Anova de un factor**
Ordinal - Numérica	Comparación de medias	T-Student; Anova de un factor
Numérica - Numérica	Correlaciones y Regresiones	**r de Pearson**; Regresión lineal

Test de hipótesis: ponen a prueba una hipótesis nula (no hay asociación) y una hipótesis alternativa (sí hay asociación). Para ello se utiliza un análisis de significancia, que debe resultar igual o bajo 0,05 (95% de confianza) para rechazar la hipótesis nula.

Chi-cuadrado > 1 Hay asociación
Chi-cuadrado < 1 No hay asociación

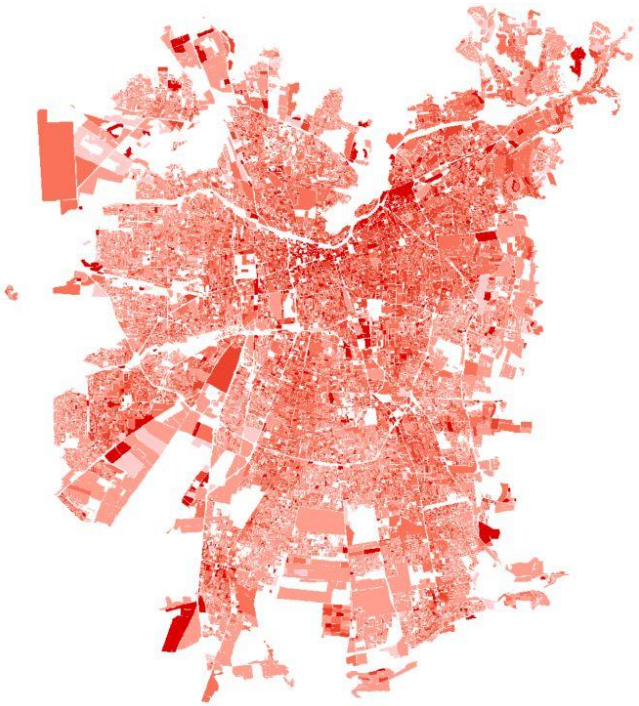
Segundo módulo

Ejercicio: Sabe leer a nivel manzana AMS

Instrucciones:

1. Abrir Redatam
2. Crear selección AMS
3. Statistical processor → Arealist

4. Table specifications → Output level: Manzent. Arrastrar variable lee.
5. Run specifications → Filtro AMS
6. Advanced → Universal filter: VIVIENDA.TIPOVIV <9 AND VIVIENDA.CONDOCUP =1
7. Run. Guardar la tabla como dbf (aunque presenta un par de problemas con símbolos como la ñ), csv o txt. Guardé como csv.
8. Abrir Arcgis.
9. Add data → Santiago2002.shp
10. Cargar la tabla de la misma manera.
11. Cerciorarme de poder hacer el Join. Al abrir la tabla nos daremos cuenta de que hay una diferencia que no permitirá hacer el join: en las manzanas el redcode está alineado a la izquierda (texto), mientras que en la tabla a la derecha (número)
    Razonamiento 1: Tengo las manzanas en texto y mi tabla en número. Pasar el número a texto. No se puede porque el csv no es editable en Arcgis.
    Razonamiento 2: Pasar de texto a número. Tampoco funciona. Si no puedo hacer esta opción, la anterior de nuevo es la única que me queda.
    Solución: Transformar la tabla a un formato editable en SIG.
12. Click derecho en la tabla → Datos → Exportar. Guardar como file geodatabase table. Pinchar "Ir a geodatabase predeterminada". Poner nombre a la tabla y guardar.
13. Ahora esta tabla permite que se le pueda añadir un campo.
    Opciones de tabla → Agregar campo → REDCODE02 Tipo texto Longitud 14
14. Click derecho arriba del nombre del nuevo campo creado. Seleccionar Calculadora de campo. → Doble click a REDCODE
15. Ahora se puede hacer el Join. Voy a las manzanas → click derecho → uniones y relaciones → unión.
16. Elegir el campo en el que se basará la unión: COD_INE02
    Elegir la tabla a unir a esta capa: Tabla_lee_exportada
    Validate join → aceptar
17. El Join es temporal, razón por la cual hay que exportar la tabla. Con esto, ya estamos en condiciones de hacer el mapa.
18. Doble click en Manzanas_join → Simbología → Cantidades → Valor: P25_1 (personas que leen). Nos va a pasar que el programa nos avisará que hay un problema con la cantidad de datos. Para solucionar eso vamos a clasificar → probando → borrar un cero al 10.000
19. Es muy obvio que las manzanas gigantes tengan más datos y por ende salgan con la mayor concentración de la categoría estudiada. Por lo tanto, la distribución no nos dice prácticamente nada. Para solucionar esto vamos de nuevo a Simbología → Cantidades → Normalización → Ponemos el total de la manzana y lo transforma en una proporción entre 0 y 1. Entonces ahora el mapa nos dirá las personas que leen en porcentajes. Para mejorar aún más el mapa podemos excluir el cero, ya que sabemos que significa no información. Clasificación → Exclusión → Población total = 0



Trabajo solo con una comuna ¿Se puede recortar de este mapa de la región?

20. COD_INE02 → Calculadora de campo → String left [COD_INE02,5]
21. Buscar código único territorial. Doble click sobre la cobertura
    Estos apuntes continúan 6:00