9. Distribución de Datos

Paula Vargas Pellicer

03/03/2022

Introducción

Vamos a seguir explorando la distribución de los datos usando un par de bases de datos:

1. La distribución de estaturas de mujeres enlistadas en el ejercito de E.U.A disponible aquí
2. La otra base de datos proviene de una publicación sobre lobos, los niveles de cortisol que presentan en condiciones de presión de caza, disponible aquí

☛ Empieza como siempre con un script nuevo e importa las bases de datos

```
setwd("tu-ruta-de-archivo")

estatura<- read.csv('ANSUR II FEMALE Public.csv', header = T,sep = ',')
lobo<- read.csv('wolf_hormone_data_for_dryad.csv', header = T,sep = ',')
```

☛ Convierte la estatura de pulgadas (in) a centímetros (cm)

Visualización de distribuciones

Gráfica

¿Qué tan alta es una mujer?

☛ Crea un histograma

La distribución de estatura es la probabiliad de observar una individua de estatura *x*

Valores con alta probabiliad: Las alturas al centro del rango (~163cm) tienen la mayor probabilidad de ser observadas

Valores con baja probabilidad: Observar una estatura mayor a 180cm es casi cero

```
length(which(estatura$cm>180))
```

Numérica

Podemos describir cualquier distribución usando estadística que describe algunos aspectos importantes de la forma de la distribución:

- La tendencia central

- La extensión (que tan ancha es la región de alta probabilidad)

- El sesgo

Tendencia central

Es el número que describe dónde ocurren la mayoría de los resultados.

Las medidas comunes son:

- La media aritmética (`mean()`)

- La mediana (`median()`)

- La moda (no hay función, créala tú)

```
median(estatura$cm, na.rm=TRUE)
```

Este número corresponde a la región de alta probabilidad en el histograma


Saca la mediana y la moda de la estatura

Extensión

Es el numero que pretende describir qué tan variables son los resultados alrededor de la tendencia central.

Las medidas comunes son:

- La desviación estándar (propagación alrededor de la media, `sd()`, usa los cuadrados)

- El rango inter-cuantil (es la diferencia entre el primer y tercer cuartil, contiene la mitad de los datos, `IQR()`)

- La desviación absoluta de la media (propagación alrededor de la media, `mad()`, usa valores absolutos)

```
ggplot(data=estatura,
        aes(x=cm)) +
     geom_boxplot() +
     labs(x='Estatura',y='Conteo') +
     theme_bw() +
     theme(axis.title = element_text(size=20),
           axis.text =  element_text(size=16))
```

☛ Calcula cada medida y identifica lo que significan en la grafica

Resúmenes numéricos poco informativos

El máximo y el mínimo tienden a ser malos descriptores porque dependen en gran medida del tamaño de la muestra. A medida que aumenta el tamaño de la muestra, es más probable que se observen valores extremos (el máximo aumentará y el mínimo disminuirá). Por esta razón, max y min no son descriptores robustos de una distribución.

```
range(estatura$cm)
```

Una medida más robusta de los extremos de una distribución son los cuantiles (por ejemplo, cuantiles del 1% y del 99%)

```
quantile(estatura$cm, probs = c(0.01,0.99))
```

☛ ¿Que significan estos números?

Distribución normal

La distribución normal (a veces llamada distribución gaussiana) es una distribución teórica mas importante en el análisis de datos. La distribución está descrita por dos parámetros:

- la media (comúnmente representada por μ)
- la desviación estándar (comúnmente representada por σ) Estos dos parámetros describen la forma de campana de la distribución normal. La distribución normal es simétrica, lo que significa que la distribución puede reflejarse sobre la media y tener el mismo aspecto.

```
# Produce un histograma
ggplot(data=estatura,
       aes(x=buttockkneelength)) +
  geom_histogram() +
  labs(x='x',
       y='Conteo') +
  theme_bw() +
  theme(axis.title = element_text(size=20),
        axis.text =  element_text(size=16))
```

☛ Saca la μ y la σ de la variable 'longitud del glúteo a la rodilla' y grafícalas en tu histograma

Distribución binomial

La distribución binomial se puede utilizar para simular la posibilidad de ver un individuo en un lugar (población), dada la probabilidad de que la especie exista en el hábitat.

```
ggplot(data=lobo,
       aes(x=Sex)) +
  geom_bar() +
  labs(x='Sexo',
       y='Conteo') +
  theme_bw() +
  theme(axis.title = element_text(size=20),
        axis.text =  element_text(size=16))
```

Aquí hay tres categorías; la tercera "U", no cuenta pues "sexo desconocido" no es un sexo.

☛ Elimina esta categoría y vuelve a graficar

☛ Identifica la variable que tiene distribución Poisson en la base de datos de los lobos.