EST-46114 Métodos Multivariados y Datos Categóricos

Sesion 03 - Inferencia en la Distribucion Gaussiana Multivariada - Parte 2

Juan Carlos Martinez-Ovando

Maestria en Ciencia de Datos

Objetivos

- Estudiaremos soluciones de problemas inferenciales comunes asociados con la clase de modelos gaussianos multidimensionales.

Prreguntas inferenciales II

Preguntas para datos tipo `swiss`, e.g.:

P1. ¿Es el tasa `Agriculture` al menos **9 puntos porcentuales** mayor a la de `Catholic`?

P2. ¿Es la tasa de `Agriculture` al menos **1.25 veces mayor** que la de `Catholic`?

P3. Otras preguntas relacionadas con la media y dispersion. . . .

El diccionario de datos `swiss` esta disponible en `https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/swiss.html`.

P1. Agriculture vs Catholic

La pregunta en cuestion puede plantearse de dos formas:
P1.a. Directamente sobre los valores de la tasa, i.e.

$$X_{Agriculture} > X_{Catholic} + 9.$$

referido a comunidades suizas, en general.
P1.b. Referidos a los valores de la tasa de una comunidad promedio, i.e.

$$\mu_{Agriculture} > \mu_{Catholic} + 9.$$

P1.b. Solucion

- La solucion para `P1.b.` es la mas sencilla, pues se refiere a los `parametros` del modelo y no a las `variables aleatorias`.
- Siendo referido a los parametros, puede a su vez resolverse de manera `lineal` o `no lineal` (refiriendo por linealidad a la propiedad y operacion sobre los parametros).

P1.b. Solucion lineal

La `solucion lineal` puede plantearse como un problema de decision bajo incertidumbre, considerando dos posibles hipotesis:

H_0: Hipotesis 0

$$\mu_{Agriculture} - \mu_{Catholic} > 9.$$

H_1: Hipotesis 1

$$\mu_{Agriculture} - \mu_{Catholic} \leq 9.$$

P1.b. Solucion lineal

Elementos del problema de decision:

1. Espacio de decisiones
   $\mathcal{A} = \mathbb{R} \times \mathbb{R} = \{(\mu_{Agriculture}, \mu_{Catholic}) : \mu_{Agriculture}, \mu_{Catholic}$ son reales$\}$.
2. Espacio de incertidumbre:
   $\Theta = \mathbb{R} \times \mathbb{R} = \{(\mu_{Agriculture}, \mu_{Catholic}) : \mu_{Agriculture}, \mu_{Catholic}$ son reales$\}$.
3. Funcion de utilidad: $U : \mathcal{A} \times \Theta \to \{0, 1\}$, siendo $U = 1$ cuando al `eleccion/decision` dle parametro coincide con el `verdadero valor del paraemto`, con $U = 0$ en el caso contrario.
4. Cuantificacion de incertidumbre: $\mathbb{P}(\mu_{Agriculture}, \mu_{Catholic} | \text{datos}, \text{complemento})$, probabilidad adecuada a la informacion de los datos.

Noten que el resto de los parametros no esta considerdo en las hipotesis, por lo que no es referido en los elementos del problema de decision.

P1.b. Solucion lineal

Noten que las hipotesis H_0 y H_1 inducen una particion sobre $\Theta = \Theta_0 \cup \Theta_1$, con

$$\Theta_0 = \{(\mu_{Agriculture}, \mu_{Catholic}) : \mu_{Agriculture} - \mu_{Catholic} > 0\},$$

y

$$\Theta_1 = \Theta \setminus \Theta_0.$$

Regla de decision:
La regla de decision consistira en elegir la hipotesis H con mayor probabilidad, i.e.

$$H^* = \arg \max_{H \in \{H_0, H_1\}} \mathbb{P}(H|\text{datos}),$$

siendo

$$\mathbb{P}(H_j|\text{datos}) = \mathbb{P}((\mu_{Agriculture}, \mu_{Catholic}) \in \Theta_j|\text{datos}).$$

Aprendizaje bayesiano

Retomemos de la sesion anterior los resultados del aprendizaje bayesiano de la clase de modelos gaussianos con los datos `swiss` usando la funcion `gaussian.posterior`:
El resultado es un objeto lista, `output`, con cuatro objetos

- `output[[1]]` - parametro $\boldsymbol{\mu}_n$
- `output[[2]]` - parametro s_n
- `output[[3]]` - parametro a_n
- `output[[4]]` - parametro $\boldsymbol{B}_n$

Aprendizaje bayesiano

```
output[{1}]
```

```
## [[1]]
##                          [,1]
## Fertility          68.68125
## Agriculture        49.60417
## Examination        16.14583
## Education          10.75000
## Catholic           40.28667
## Infant.Mortality 19.52708
```

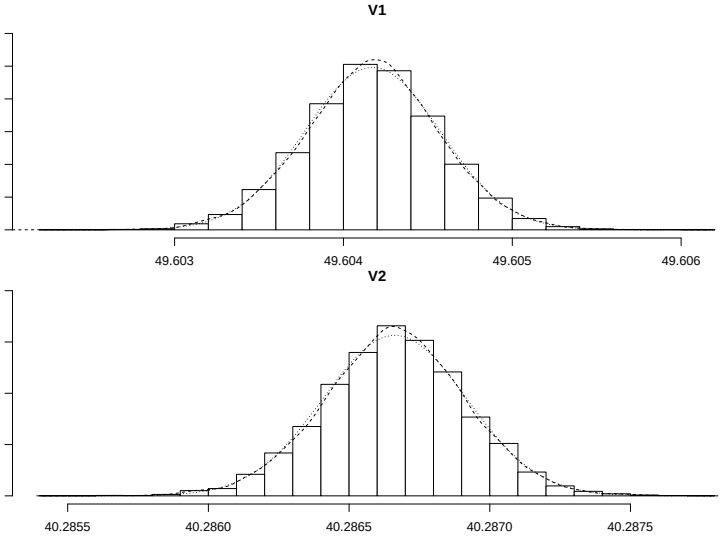
En primera instancia, pareciera ser que H_0 se cumple con estos datos. Sin embargo, $\mu_{Agriculture,n}$ y $\mu_{Catholic,n}$ son estimadores de cantidades desconocidas, por lo que la relacion referida a H_0 dependendera de la variabilidad epistemica de estos parametros, en funcion de a_n, s_n y B_n.

Incertidumbre epistemica

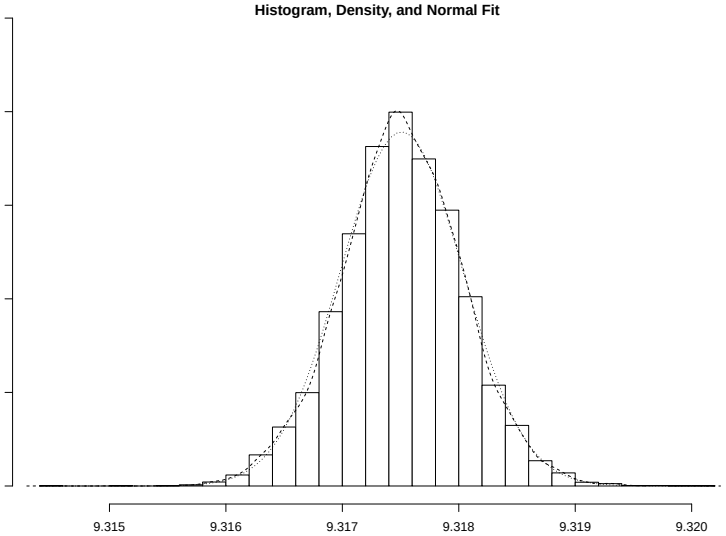
La **incertidumbre epistemica** se refiere a la cuantificacion del desconocimiento acerca de $(\mu_{Agriculture,n}, \mu_{Catholic,n})$, en este caso. tal deconocimiento, marginal, esta referido a:

- Distribuciones marginales de $\mu_{Agriculture}$ y $\mu_{Catholic}$
- Interaccion / comovimiemto / covarianza entre $\mu_{Agriculture}$ y $\mu_{Catholic}$

Visualizacion I



Visualizacion II



Regla de decision (solucion lineal)

La par probabilidades para H_0 y H_1 (aproximadas por Monte Carlo) son

```
prob_lin_h0 <- length(which(mu.sim.2[,1]-mu.sim.2[,2] > 9))/M
prob_lin_h0
```

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

y

```
prob_lin_h1 <- length(which(mu.sim.2[,1]-mu.sim.2[,2] <= 9))/M
prob_lin_h1
```

```
## [1] 0
```

En este caso, la evidencia es contundente. **Consideren que este NO es el caso, en general.**

P1.b. Solucion lineal exacta

La solucion del problema, que hemos resulto via `simulacion` se pudo haber obtenido analiticamente empleando el resultado de cerradura bajo linealidad, considerando la aplicacion del vector

$$\boldsymbol{c}_{Hipotesis} = (0, 1, 0, 0, -1, 0),$$

a

$$\boldsymbol{\mu},$$

considernado que

$$(\boldsymbol{\mu}, \boldsymbol{\Lambda}) \sim \text{N-Wi}(\boldsymbol{\mu}, \boldsymbol{\Lambda} | \boldsymbol{m}_n, s_n, a_n, \boldsymbol{B}_n).$$

- Se sigue que $\boldsymbol{c}'_{Hipotesis}\boldsymbol{\mu}$ sigue una distribucion t-univariada.
- Los datos que simulamos, son una muestra 'iid' de esta distribucion marginal.
- Solo problemas asociados con tranformaciones afines de $\boldsymbol{X}$ o $\boldsymbol{\mu}$ pueden resolverse de manera analitica cerrada. En general, sera util recurrir a herramientas de `simulacion estocastica`.

P2. Solucion no lineal

Ahora, la `solucion no lineal` a la pregunta P2 puede plantearse como un problema de decision bajo incertidumbre, considerando dos posibles hipotesis:

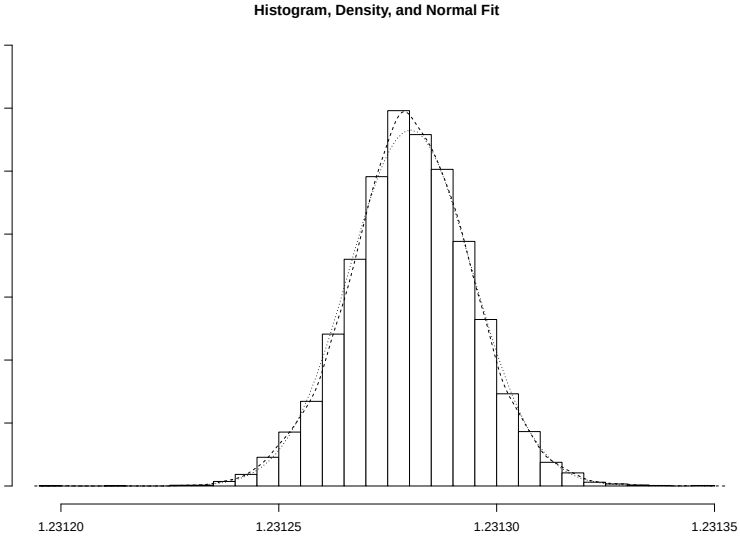
H_0: Hipotesis 0

$$\mu_{Agriculture}/\mu_{Catholic} > 1.25.$$

H_1: Hipotesis 1

$$\mu_{Agriculture}/\mu_{Catholic} \leq 1.25.$$

Visualizacion III



Regla de decision (solucion lineal)

La par probabilidades para H_0 y H_1 (aproximadas por Monte Carlo) son

```
prob_nolin_h0 <- length(which(mu.sim.2[,1]/mu.sim.2[,2] > 1.25))/M
prob_nolin_h0
```

```
## [1] 0
```

y

```
prob_nolin_h1 <- length(which(mu.sim.2[,1]/mu.sim.2[,2] <= 1.25))/M
prob_nolin_h1
```

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

En este caso, la evidencia es contundente. **Consideren que este NO es el caso, en general.**

Ejercicio

Desarrollen las soluciones al problema `lineal` y `no lineal` deferido ahora a las variables

$$X_{Agriculture} \text{ y } X_{Catholic},$$

en lugar de referirlo a $\mu_{Agriculture}$ y $\mu_{Catholic}$.
Desarrollen las soluciones via simulacion, considerando lo siguiente:

$$\begin{aligned} \boldsymbol{\Lambda}|\text{datos} &\sim \text{Wi}(\boldsymbol{\Lambda}|\boldsymbol{a_n}, \boldsymbol{B_n}) \\ \boldsymbol{\mu}|\boldsymbol{\Lambda}, \text{datos} &\sim \text{N}(\boldsymbol{\mu}|\boldsymbol{m}_n, s_n\boldsymbol{\Lambda}) \\ \boldsymbol{X}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \text{datos} &\sim \text{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}). \end{aligned}$$

En la siguiente sesion

Revisaremos como hacer:

1. Descomposion singular de matrices de covarianzas
2. Analisis de componentes principales
3. Analisis inferencial de componentes principales

Lecturas complementarias

- Hothorn et al (2018) "On multivariate t and Gauss probabilities in R". `est46114_s04_suplemento1.pdf`
- Martinez-Ovando (2016) "Paradigma Bayesiano de Inferencia (Resumen de Teoria)", *Notas de clase EST-46114*. (De momento, no presten antencion a la descripcion de los metodos de simulacion transdimensionales.) `est46114_s03_suplemento2.pdf`
- Press (2005). "Applied multivariate analysis, using Bayesian and frequentist methods of inference." Dover Pub. (Capitulo 4.)

Table of Contents