Clase 14 de Bioestadística

Estadística: estimación de la esperanza y la varianza

Ernesto Mordecki

CMAT, Facultad de Ciencias, Universidad de la República.

Uruguay

Contenidos de la clase

Probabilidad y Estadística

- La probabilidad y la estadística son disciplinas hermanas
- Comparten un mismo marco teórico matemático:
  - un espacio de sucesos Ω
  - un modelo de probabilidad P
- Se diferencian en las pregunas que formula cada una:
  - La probabilidad conoce P y calcula a partir de ella
  - La estadística conoce valores $X_1, \ldots, X_n$ y se ocupa de estimar[1] P.
  - La estimación de la probabilidad se designa $\widehat{P}$.

[1] Se trata de dar una buena aproximación del modelo

Ejemplo 1: Modelo normal

Consideramos un ejemplo clásico

- Suponemos el modelo gaussiano o normal donde desconocemos la media μ y la varianza σ^2
- Observamos (conocemos) n valores numéricos de variables aleatorias $X_1, \ldots, X_n$ que corresponden al modelo incóngnita
- ¿Cómo estimamos los parámetros μ y σ^2 a partir de los datos $X_1, \ldots, X_n$?

Ejemplo 2: Estimación de una probabilidad

Consideramos otro ejemplo clásico

- Suponemos el modelo Bernuolli donde desconocemos la probabilidad p
- Observamos (conocemos) n valores numéricos de variables aleatorias $X_1, \ldots, X_n$ que corresponden al modelo incóngnita, que corresponden a una tira de ceros (fracasos) y unos (éxitos)
- ¿Cómo estimamos el parámetro p a partir de esa tira de ceros y unos?

Terminología estadísitica

- El conunto Ω se llama espacio muestral. En probablidad se llama espacio de sucesos
- Los valores $X_1, \ldots, X_n$ se llaman muestra aleatoria simple, se resume m.a.s[2]
- Si el parámetro desconocido se designa θ, la aproximación que hacemos se designa $\widehat{\theta}$.

[2] Es lo que en probabilidad se llama v.a.i.i.d.

Estimación de la esperanza

- La base de la estimación es la ley fuerte de los grandes números (LFGN):
- Sean $X_1, \ldots, X_n$ una m.a.s. de v.a. con esperanza desconocida $\mu = E(X_1)$.
- La LFGN establece que:

$$\bar{X}_n = \frac{X_1 + \cdots + X_n}{n} \to E(X_1)$$

- Si tenemos un modelo con una probabilidad P con $\mu = E(X_1)$ desconocido, podemos estimar μ a traves del promedio de los datos

$$\widehat{\mu} = \overline{X}_n$$

- La LFGN nos dice entonces que

$$\widehat{\mu} \to \mu,$$

que se denomina consistencia del estimador.

Estimación de una probabilidad

- Supongamos un modelo de Bernoulli con probabilidad p desconocida.
- Tenemos una m.a.s. $X_1, \ldots, X_n$. Se verifica

$$E(X_1) = p$$

- Entonces se reduce al caso anterior: ¡la probabilidad es la esperanza!

- Entonces, el estimador de la probabilidad es

$$\widehat{p} = \frac{X_1 + \cdots + X_n}{n} = \frac{\text{cantidad de éxitos}}{n}$$

- El estimador también se denomina frecuencia observada de la cantidad de éxitos

- Además, la LFGN nos dice que el estimador es consistente, es decir

$$\widehat{p} \to p.$$

Estimación de la varianza con media conocida

- Tenemos ahora una m.a.s. $X_1, \ldots, X_n$, suponemos que $\mu = E(X_1)$ es conocida, pero $\sigma^2 = var(X_1)$ es desconocida
- Queremos entonces estimar σ^2.
- Recordamos para eso que σ^2 es en realidad una esperanza, de la variable aleatoria

$$Y = (X - E(X))^2 = (X - \mu)^2$$

- Producimos entonces una m.a.s. de Y, mediante

$$Y_i = (X_i - \mu)^2$$

- Estimamos entonces la varianza mediante

$$\widehat{\sigma^2} = \frac{(X_1 - \mu)^2 + \cdots + (X_n - \mu)^2}{n}$$

- Aplicando la LFGN sabemos que

$$\widehat{\sigma^2} = \frac{(X_1 - \mu)^2 + \cdots + (X_n - \mu)^2}{n} \to E(X - \mu)^2 = \sigma^2,$$

es decir, tenemos consistencia.

Estimación de la varianza con media **desconocida**

- La situación más frecuente es cuando no conocemos ni la media μ ni la varianza σ^2 de la variable que queremos estimar.
- Por lo tanto no conocemos μ para ponerlo en la fórmula anterior.
- La idea es entonces primero estimar μ, y luego utilizarlo para estimar la varianza.
- El resultado de este procedimiento es

$$\widehat{\sigma^2} = \frac{(X_1 - \widehat{\mu})^2 + \cdots + (X_n - \widehat{\mu})^2}{n}$$

- En este caso las variables

$$Y_i = (X_i - \widehat{\mu})^2$$

no conforman una muestra aleatoria simple, porque no son independientes.

- Se puede probar sin embargo que

$$\widehat{\sigma^2} = \frac{(X_1 - \widehat{\mu})^2 + \cdots + (X_n - \widehat{\mu})^2}{n} \to \sigma^2,$$

también tenemos consistencia.

Estimación del desvío estándar

- Una vez estimada la varianza, para estimar el desvío estándar se toma la raíz:

$$\widehat{\sigma_X} = \sqrt{\widehat{\sigma^2}} = \sqrt{\frac{(X_1 - \widehat{\mu})^2 + \cdots + (X_n - \widehat{\mu})^2}{n}}$$

$$= \sqrt{\frac{1}{n}\sum_{k=1}^{n}(X_k - \widehat{\mu})^2}$$

Correlación y su estimación

- En el caso en que dos variables no son independientes, decimos que son dependientes
- Una medida de su dependencia es la covarianza, que según vimos se define como

$$cov(X, Y) = E\left[(X - E(X))(Y - E(Y))\right]$$

- Con el objetivo de obtener una cantidad adimensionada, definimos la correlación

$$\rho(X, Y) = \frac{E\left[(X - E(X))(Y - E(Y))\right]}{\sigma_X \sigma_Y}$$

donde σ_X y σ_Y son los desvíos estándar de X e Y respectivamente.

Propiedades

- Se puede demostrar que la correlación es una cantidad que se encuentra entre -1 y 1, es decir

$$-1 \leq \rho(X, Y) \leq 1.$$

- La correlación positiva indica que a valores mayores de una se obtienen valores mayores de la otra,
- La correlación negativa indica lo contrario: valores mayores que la esperanza de una corresponden a valores menores que la esperanza de la otra
- Cuanto más cerca de uno (o de -1) sea el valor de la correlación, mas pronunciado será este fenómeno.

Estimación

- Para estimar la covarianza suponemos conocer una muestra aleatoria simpe del vector aleatorio (X, Y), es decir, v.a.s

$$(X_1, Y_1), \ldots, (X_n, Y_n)$$

- Aquí se supone que las parejas son independientes entre sí, pero no tiene que serlo cada X_k de su compañera Y_k

Estimación con medias y desvíos conocidos

- Supongamos primero que conocemos $\mu_X = E(X_1)$, $\mu_Y = E(Y_1)$ así como las varianzas $\sigma_X^2 = var(X_1)$ y $\sigma_Y = var(Y_1)$
- En ese caso la correación es la esperanza de las variables

$$Z = \frac{(X - \mu_X)(Y - \mu_Y)}{\sigma_X \sigma_Y}$$

- Entonces conseguimos un estimaror consistente aplicando la LFGN:

$$\widehat{\rho}(X, Y) = \frac{1}{n} \sum_{k=1}^{n} \frac{(X_k - \mu_X)(Y_k - \mu_Y)}{\sigma_X \sigma_Y}$$

Estimación: caso general

- En la situación se desconocen las esperanzas y las varianzas.
- La estimación entonces comienza estimando las esperanzas μ_X μ_Y y los desvíos σ_X , σ_Y.
- Primero estimamos las esperanzas:

$$\widehat{\mu_X} = \overline{X}_n, \qquad \widehat{\mu_Y} = \overline{Y}_n$$

- y luego los desvíos estándar

$$\widehat{\sigma_X} = \sqrt{\frac{1}{n}\sum_{k=1}^{n}(X_k - \widehat{\mu_X})^2}$$

$$\widehat{\sigma_Y} = \sqrt{\frac{1}{n}\sum_{k=1}^{n}(Y_k - \widehat{\mu_Y})^2}$$

- Finalmente construímos el estimador de la correlación:

$$\widehat{\rho}(X, Y) = \frac{1}{n}\sum_{k=1}^{n}\frac{(X_k - \widehat{\mu_X})(Y_k - \widehat{\mu_Y})}{\widehat{\sigma_X}\widehat{\sigma_Y}}$$

- El estimador de la correlación sirve para saber si dos variables son dependientes (¿cómo?)