Introducción a la Inferencia basada en modelos

Muestreo y Planificación de Encuestas.

Primer Semestre 2023

Tipos de Inferencia*

- **Basada en el diseño:** la aleatoriedad proviene del mecanismo de selección de la muestra. Los datos observados son considerados como fijos.

- **Basada en modelos:** las observaciones tienen una distribución de probabilidad asumida, reflejada en un modelo superpoblacional.

La inferencia basada en el diseño tiene propiedades asintóticas, por lo cual, si esto no se cumple, los supuestos que la sostienen comienzan a fallar. Esto motiva cambiar de paradigma en algunas situaciones, y por eso la importancia de presentar los conceptos básicos de la inferencia basada en modelos.

*La presentación se encuentra basada en el libro *An Introduction to Model-Based Survey Sampling* de Chambers y Clark, 2012

Modelo Superpoblacional

Para especificar las propiedades estadísticas de los valores poblacionales de la variable de interés se propone un modelo, el cual se denomina como **superpoblacional**.

Ejemplo

Dada y_k y una variable auxiliar x_k, con $k = 1, \ldots, N$, se puede asumir que la población fue generada por el modelo

$$E(y_k|x_k) = \beta_0 + \beta_1 x_k$$

Los parámetros superpoblacionales son β_0 y β_1.

La teoría estadística provee varios métodos para estimar de forma eficiente β_0 y β_1 (Modelos Lineales). Éstos asumen que no existe relación entre los valores generados por el modelo y el método utilizado para seleccionar la muestra.

Esto no sucede cuando se trabaja con muestras reales e ignorar el diseño muestral puede inducir a sesgos en la estimación de β_0 y β_1.

Diseños No Informativos

Sea θ el parámetro de interés, definido por los valores de y.

- X_U es la información auxiliar a nivel de la población.
- Y_U es la variable de interés a nivel de la población.
- Y_s son los valores de Y observados en la muestra s.

Un diseño es **no informativo** para realizar inferencias sobre θ dada X_U si la distribución conjunta de $Y_s|X_U$ es la misma que $Y_U|X_U$.

Es decir, el modelo superpoblacional asumido ajusta tanto en la muestra como en la población.

El método de muestreo sólo interviene en la inferencia sobre los parámetros del modelo, la muestra **no contiene información "extra"** acerca de la variable y y los parámetros de interés.

Como consecuencia la muestra s debería ser tratada como **dada** cuando se realizan inferencias sobre los parámetros de la distribución conjunta del vector Y, dada la información auxiliar.

LOS DISEÑOS PROBABILÍSTICOS SON NO INFORMATIVOS

Pueden depender de los valores de una variable Z, pero una vez que condicionamos sobre los valores de Z, el resultado del proceso de selección no depende de los valores de Y.

Diseños No Informativos

¿Por qué es importante que el diseño sea no informativo en el contexto de la inferencia basada en modelos?

Porque vamos a realizar inferencias con el modelo en el conjunto no observado $r = U - s$.

Si el diseño es no informativo podemos decir que el modelo funciona tanto en s como en $U - s$.

Diseños No Informativos

¿Qué pasa si el muestreo es informativo?

- Puedo tener información sobre cómo "informa" el diseño, y lo incluyo en el modelo (poco frecuente).
- Se utilizan métodos robustos para realizar inferencias. Sólo funciona si la distribución de $Y_s|Z$ y $Y_U|Z$ no son muy diferentes. Si hay una total desconexión entre estas dos distribuciones, no se resuelve con una estimación robusta.

Diseños No Informativos

¿Cómo evitamos que el diseño sea informativo?

- Eligiendo la muestra con un diseño probabilístico.
- Minimizando la tasa de no respuesta.

El enfoque basado en modelos

El parámetro de interés es $t_y = \sum_U y_k$ y su estimador es $\hat{t}_y$. Tenemos que:

- Los y_k son aleatorios (generados por el modelos superpoblacional).
- Las esperanzas y las varianzas son condicionales a la muestra s (se toma como fija).
- t_y es la suma de variables aleatorias, por lo tanto es una variable aleatoria.
  Estimar t_y es equivalente a *predecir* el valor de esta variable aleatoria utilizando los valores observados en s.
- Los "predictores" $\hat{t}_y$ de t_y son funciones de los $y_k \in s$ y de los z_k. $\hat{t}_y$ también es una variable aleatoria.
- Los parámetros del modelo β_0 y β_1 son desconocidos y se asumen como fijos.

El enfoque basado en modelos

Notemos que tanto t_y como $\hat{t}_y$ son realizaciones de variables aleatorias, cuya distribución conjunta es determinada por dos procesos:

- El modelo superpoblacional.
- El proceso por el cual se elige la muestra s ("posiblemente aletoria, posiblemente no").

Para en análisis de los predictores se asume que el diseño es no informativo.

El enfoque basado en modelos

Queremos que t_y y $\hat{t}_y$ sean cercanos. Nos vamos a enfocar en:

$$E(\hat{t}_y - t_y)$$

y

$$Var(\hat{t}_y - t_y)$$

La esperanza y la varianza... ¿respecto a qué se calculan?

** Notar que en la notación Z es ahora la variable auxiliar del modelo

El enfoque basado en modelos

Queremos que t_y y $\hat{t}_y$ sean cercanos. Nos vamos a enfocar en:

$$E(\hat{t}_y - t_y)$$

y

$$Var(\hat{t}_y - t_y)$$

La esperanza y la varianza... ¿respecto a qué se calculan?

Respecto al modelo dada Z**

**Notar que en la notación Z es ahora la variable auxiliar del modelo

El enfoque basado en modelos

Notemos que:

$$t_y = \sum_s y_k + \sum_r y_k = t_{ys} + t_{yr}$$

El problema es entonces predecir t_{yr}. Tenemos entonces que:

$$\hat{t}_y = t_{ys} + \hat{t}_{yr}$$

Surgen dos preguntas:

- Dado el modelo asumido, ¿cuál es el "mejor" predictor $\hat{t}_{yr}$ de t_{yr}?
- Dado el modelo asumido y el "mejor" predictor, ¿cuál es la mejor forma de seleccionar la muestra s con el fin de minimizar el error $\hat{t}_y - t_y = \hat{t}_{yr} - t_{yr}$?

La respuesta depende de nuestra definición de "mejor" y de "minimizar":

Notemos que:

$$t_y = \sum_s y_k + \sum_r y_k = t_{ys} + t_{yr}$$

El problema es entonces predecir t_{yr}. Tenemos entonces que:

$$\hat{t}_y = t_{ys} + \hat{t}_{yr}$$

Surgen dos preguntas:

- Dado el modelo asumido, ¿cuál es el "mejor" predictor $\hat{t}_{yr}$ de t_{yr}?
- Dado el modelo asumido y el "mejor" predictor, ¿cuál es la mejor forma de seleccionar la muestra s con el fin de minimizar el error $\hat{t}_y - t_y = \hat{t}_{yr} - t_{yr}$?

La respuesta depende de nuestra definición de "mejor" y de "minimizar":

i) $\hat{t}_y$ pertenece a una clase de predictores "aceptables" de t_y.

ii) $\hat{t}_y$ tiene el menor valor de $E(\hat{t}_y - t_y)^2$ dentro de la clase de predictores "aceptables", dada la muestra s.

Predicción óptima

Las propiedades estadísticas de $\hat{t}_y$ se encuentran definidas por la distribución del error $\hat{t}_y - t_y$ bajo el modelo superpoblacional.

Sesgo de predicción:

$$E(\hat{t}_y - t_y)$$

Varianza de la predicción:

$$Var(\hat{t}_y - t_y)$$

Error Cuadrático Medio (MSE)

$$E(\hat{t}_y - t_y)^2 = Var(\hat{t}_y - t_y) + \{E(\hat{t}_y - t_y)\}^2$$

Predicción óptima

El primer paso es identificar un predictor óptimo para t_{yr} dada s.

Repaso

El predictor que minimiza el MSE de una variable aleatoria W dada una variable aleatoria V es:

$$E(W|V)$$

Lo aplicamos con $W = t_y$ y $V = s$:

Resultado

El predictor óptimo de t_y es:

$$t_y^* = E(t_y|y_k \in s; z_k, k = 1, ..., N) = t_{ys} + E(t_{yr}|y_k \in s; z_k, k = 1, ..., N)$$

Predicción óptima

Si el modelo es $E(y_k|z_k) = \beta_0 + \beta_1 z_k$, tenemos que:

$$t_y^* = t_{ys} + \sum_r \beta_0 + \beta_1 z_k$$

Entonces, se estiman β_0 y β_1, y obtenemos el estimador *Empirical Best* (EB) de t_y:

$$t_y^{EB} = t_{ys} + \sum_r \hat{\beta}_0 + \hat{\beta}_1 z_k = t_{ys} + \sum_r \hat{y}_k$$

Notemos que es clave que el diseño sea NO INFORMATIVO.