# Capítulo 12 Consideraciones al diseñar experimentos

## 12.1 Los elementos que afectan la información en una muestra

Una medida significante de la información que contiene una muestra para hacer una inferencia acerca de un parámetro poblacional es proporcionada por el ancho del intervalo de confianza que se puede construir. Recordamos que un intervalo de confianza de muestra grande de 95% para una media poblacional es

$$\overline{Y} \pm 1.96\left(\frac{\sigma}{\sqrt{n}}\right)$$

Es claro ver como este tipo de intervalos dependen de la varianza poblacional y el tamaño muestral. Si la varianza poblacional es pequeña entonces el intervalo será corto. Del mismo modo el ancho del intervalo disminuye cuando el tamaño de la muestra aumenta. Entonces estos son dos factores que afectan a la cantidad de información en una muestra relacionada con un parámetro. En este capítulo se verán distintas formas de diseñar un experimento y la relación que tienen con estos factores.

## 12.2 Diseño de experimento para aumentar la precisión

Existe distintas maneras de recolectar datos de tal manera que proporcionen más información respecto a parámetros poblaciones. No hay un solo diseño que es mejor que todo los demás. Por esta razón solo ser verán dos ejemplos para ilustrar los principios involucrados.

Considere el problema de calcular la diferencia entre un par de medias poblacionales, $\mu_1 - \mu_2$ con base en muestras aleatorias independientes. En el caso de que el experimentador tiene recursos suficientes para muestrear un total de $n$ observaciones una pregunta interesante es cuantas observaciones se deben de seleccionar de las poblaciones, es decir, $n_1$ y $n_2$ $(n_1 + n_2 = n)$, respectivamente para maximizar la información de los datos pertinentes a $\mu_1 - \mu_2$.

Si las muestras aleatorias son independientes, se puede estimar $\mu_1 - \mu_2$ con $\overline{Y}_1 - \overline{Y}_2$ que tiene error estándar

$$\sigma_{\overline{Y}_1 - \overline{Y}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Cuanto menor sea este error estándar, menor será el correspondiente error de cálculo y mayor será la cantidad de información de la muestra pertinente a la diferencias de medias. Si se asume que $\sigma_1^2 = \sigma_2^2 = \sigma^2$, entonces

$$\sigma_{\overline{Y}_1 - \overline{Y}_2} = \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Esta cantidad es mínima cuando $n_1 = n_2$. En consecuencia, la muestra contiene un máximo de información acerca de $\mu_1 - \mu_2$ cuando las $n$ unidades experimentales se dividen por igual entre los dos tratamientos.

**Demostración**

Si han de usarse $n$ observaciones para calcular $\mu_1 - \mu_2$ con base en muestras aleatorias independientes de las dos poblaciones de interés. Suponga que $n_1 + n_2 = n$.

Denote con $b$ la fracción de las $n$ observaciones asignadas a la muestra de la población 1; esto es, $n_1 = bn$ y $n_2 = (1-b)n$. Entonces,

$$V(\overline{Y}_1 - \overline{Y}_2) = \frac{\sigma_1^2}{bn} + \frac{\sigma_2^2}{(1-b)n}$$

Para hallar la fracción $b$ que minimiza esta varianza, igualamos a cero la primera derivada con respecto a $b$. Este proceso da

$$-\frac{\sigma_1^2}{n}\left(\frac{1}{b^2}\right) + \frac{\sigma_2^2}{n}\left(\frac{1}{1-b}\right)^2 = 0$$

Si despejamos $b$ tendremos,

$$b = \frac{\sigma_1}{\sigma_1 + \sigma_2} \qquad 1 - b = \frac{\sigma_2}{\sigma_1 + \sigma_2}$$

Por lo tanto, $V(\overline{Y}_1 - \overline{Y}_2)$ está minimizada cuando

$$n_1 = \left(\frac{\sigma_1}{\sigma_1 + \sigma_2}\right)n \qquad n_2 = \left(\frac{\sigma_2}{\sigma_1 + \sigma_2}\right)n$$

Es decir, cuando tamaños muestrales se asignan de manera proporcional a tamaños de desviaciones estándar. Observe que $n_1 = n/2 = n_2$ si $\sigma_1 = \sigma_2$.

En el segundo ejemplo se considera un modelo de regresión lineal, particularmente el parámetro $\beta_1$. La desviación estándar de el estimador $\hat{\beta}_1$ es,

$$\sigma_{\hat{\beta}_1} = \frac{\sigma}{\sqrt{S_{xx}}} = \frac{\sigma}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}}$$

Cuanto mayor sea $S_{xx}$, la suma de cuadrados de las desviaciones de $x_1, x_2, \ldots, x_n$ alrededor de su media, menor será la desviación estándar de $\hat{\beta}_1$. En otras palabras, se obtendrá un mejor estimador para la pendiente si los valores de $x$ están más dispersos. Esta dispersión varia de caso en caso, por lo general su efectividad se basa en el rango que tengan la variable $x$. Además, si los valores de $x$ están más dispersos, la validez de la suposición de que la varianza del término de error $\varepsilon$ no depende del valor de la variable independientes $x$.

## 12.3 El experimento de las observaciones pareadas

En muchos experimentos las muestras son pareadas más que independientes. Una situación que ocurre comúnmente es aquella donde se hacen observaciones repetidas en la misma unidad de muestreo. La comparación de dos poblaciones con base en datos pareados puede ser un diseño experimental muy eficaz que puede controlar fuentes extrañas de variabilidad y resultar en la disminución del error estándar del estimador $\overline{Y}_1 - \overline{Y}_2$ para diferencia en las medias poblacionales $\mu_1 - \mu_2$. Con $(Y_{1i}, Y_{2i})$, para $i = 1, 2, \ldots, n$, denote una muestra aleatoria de observaciones pareadas. Suponga que,

$$E(Y_{1i}) = \mu_1 \qquad Var(Y_{1i}) = \sigma_1^2 \qquad E(Y_{2i}) = \mu_2$$
$$Var(Y_{2i}) = \sigma_2^2 \qquad Cov(Y_{1i}, Y_{2i}) = \rho\sigma_1\sigma_2$$

donde $\rho$ es el coeficiente de correlación común de las variables dentro de cada par. Definimos con $D_i = Y_{1i} - Y_{2i}$, para $i = 1, 2, \ldots, n$, las diferencias entre las obsrevaciones dentro de cada par. Debidoa que los pares de observaciones se supusieron independientes y distribuidos idénticamente, los valores $D_i$, para $i = 1, 2, \ldots, n$, son independientes y distribuidos idénticamente. Al usar el Teorema 5.12 se obtiene lo siguiente,

$$\begin{aligned}\mu_D &= E(D_i) = E(Y_{1i}) - E(Y_{2i}) = \mu_1 - \mu_2\\ \sigma_D^2 &= Var(D_i) = Var(Y_{1i}) + Var(Y_{2i}) - 2Cov(Y_{1i}, Y_{2i})\\ &= \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2\end{aligned}$$

De esto ultimo, se deduce que un estimador natural para $\mu_1 - \mu_2$ es el promedio de las diferencias $\overline{D} = \overline{Y}_1 - \overline{Y}_2$, con esto se obtiene,

$$E(\overline{D}) = \mu_D = \mu_1 - \mu_2$$
$$\sigma_{\overline{D}}^2 = Var(\overline{D}) = \frac{\sigma_D^2}{n} = \frac{1}{n}\left[\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2\right]$$

Si la información se había obtenido de un experimento con muestras independientes y $n_1 = n_2 = n$, se obtiene lo siguiente,

$$E(\overline{Y}_1 - \overline{Y}_2) = \mu_1 - \mu_2$$
$$\sigma_{\overline{Y}_1 - \overline{Y}_2}^2 = \frac{1}{n}\left[\sigma_1^2 + \sigma_2^2\right]$$

Si es razonable creer que en los pares $(Y_{1i}, Y_{2i})$, para $i = 1, 2, \ldots, n$, los valores de $Y_{1i}$ y $Y_{2i}$ tenderán a aumentar o disminuir juntos, entonces un análisis de las expresiones anteriores para $\sigma_{\overline{D}}^2$ del experimento de pares acoplados y $\sigma_{\overline{Y}_1 - \overline{Y}_2}^2$ del experimento de muestras independientes demuestra que el experimento de observaciones pareadas proporciona un estimador con menor varianza que el de muestras independientes.

Debido a que parear muestras hace dependientes las observaciones dentro de cada par, no podemos usar los método que previamente

se desarrollaron para comparar poblaciones con base en muestras independientes entre sí. Este análisis de las observaciones pareadas utilizan las $n$ diferencias pareadas, $D_i$, para $i = 1, 2, \ldots, n$. Las inferencias con respecto a las diferencias en las medias $\mu_1 - \mu_2$ se construyen haciendo inferencias respecto a la media de las diferencias $\mu_D$. Defina

$$\overline{D} = \frac{1}{n}\sum_{i=1}^{n} D_i \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \overline{D})^2$$

y utilice el procedimiento apropiado de una muestra para completar la inferencia. Si se tiene una muestra lo suficientemente grande se pueden desarrollar métodos inferenciales de muestra grande. Si no se tiene una muestra lo suficientemente grande entonces es razonable asumir que las *diferencias están distribuidas normalmente en forma aproximada* y se pueden ocupar métodos inferenciales basados en la distribución $t$. Al utilizar las diferencias se tienen menos restricciones de las suposiciones de las poblaciones originales. Lo que nos ayuda en obtener inferencias de estas de manera sencilla.

## 12.4 Algunos diseños experimentales elementales

En esta sección presentamos consideraciones generales asociadas con diseñar experimentos. En especial consideramos extensiones de las metodologías de muestras independientes y de observaciones pareadas cuando el objetivo es comparar medias de más de dos poblaciones. A continuación presentaremos los componentes básicos del diseño de un experimento.

**Definición 12.1**

Las *unidades experimentales* son los objetos sobre los que se toman mediciones.

**Definición 12.2**

Los *factores* son variables controladas completamente por el experimentador. El nivel de intensidad (Subcategoría distinta) de un factor es su *nivel*.

Considere un experimento realizado para investigar el efecto de varias cantidades de nitrógeno y fosfato en la producción de una variedad de maíz. Una *unidad experimental* sería una superficie especificada, por ejemplo 1 acre, de maíz. Un *tratamiento* será un número fijo de libras de nitrógeno $x_1$ y de fosfato $x_2$ aplicados a un acre determinado de maíz. Por ejemplo, un tratamiento podría ser usar $x_1 = 100$ libras de nitrógeno por acre y $x_2 = 200$ libras de fosfato. Un segundo tratamiento podría corresponder a $x_1 = 150$ y $x_2 = 100$. Observe que el experimentador podría usar diferentes cantidades $(x_1, x_2)$ de nitrógeno y fosfato y que cada *combinación* representaría un tratamiento diferente.

**Definición 12.3**

Un *tratamiento* es una combinación específica de niveles de factor.

**Definición 12.4**

Un *diseño completamente aleatorizado* para comparar $k$ tratamientos es aquel en el que un grupo de $n$ unidades experimentales relativamente homogéneas se dividen al azar en $k$ subgrupos de tamaños $n_1, n_2, \ldots, n_k$ (donde $n_1 + n_2 + \cdots + n_k = n$). Todas las unidades experimentales de cada subgrupo reciben el mismo tratamiento, con cada tratamiento aplicado a exactamente un subgrupo.

Asociada a cada tratamiento está una población (a veces conceptual= consistente en todas las observaciones que habrían resultado si el tratamiento se aplicara en forma repetida. Las observaciones obtenidas de un diseño completamente aleatorizado se ven de manera típica como *muestras aleatorias independientes* tomadas de las poblaciones correspondientes a cada uno de los tratamientos.

**Definición 12.5**

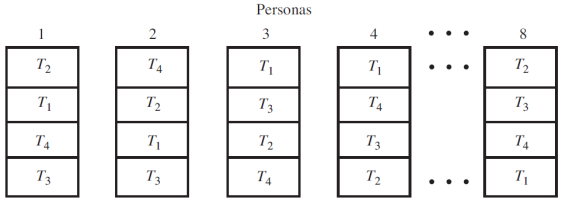
Un *diseño biunívoco* para comparar $k$ poblaciones es un arreglo en el que se obtienen muestras aleatorias de cada una de las poblaciones de interés.

En esta forma, un diseño biunívoco, ya sea que corresponda a datos obtenidos con el uso de un diseño completamente aleatorizado o al tomar muestras independientes de cada una de varias poblaciones existentes, es la extensión de los experimentos muestrales independientes que consideramos en los Capítulos 8 y 10. El objetivo de este diseño es idéntico al del diseño de observaciones pareadas, es decir, eliminar fuentes no deseadas de variabilidad que podrían entrar de manera subrepticia en las observaciones de nuestro experimento.

**Definición 12.6**

Un *diseño aleatorizado en bloque* contiene $b$ bloques y $k$ tratamientos consistentes de $b$ bloques de $k$ unidades experimentales cada uno. Los tratamiento se asignan de manera aleatoria a las unidades de cada bloque, con cada tratamiento apareciendo exactamente una vez en cada bloque.

Al compara tratamientos dentro de bloques de material experimental relativamente homogéneos, se puede usar para eliminar variación de un bloque a otro cuando se comparan tratamientos.




El diseño de bloques en dos direcciones se puede lograr con el uso de un diseño de cuadro latino. Este tipo de diseños se ven en la siguiente figura,