UNIVERSIDAD DE LA REPÚBLICA

Facultad de Ciencias Económicas y de Administración

Licenciatura en Estadística

Vinculación entre infecciones parasitarias intestinales y estado nutricional en escolares de la escuela 317 (Centro Comunal Zonal 6 de Montevideo)

Federico Alvarez - Fernando Massa

Tutores:
Ramón Alvarez
Laura Nalbarte

Montevideo, Marzo 2012.

UNIVERSIDAD DE LA REPÚBLICA

FACULTAD DE CIENCIAS ECONOMICAS Y ADMINISTRACION

El tribunal docente integrado por los abajo firmantes aprueba el trabajo de Pasantía:

Vinculación entre infecciones parasitarias intestinales y estado nutricional en escolares de la escuela 317 (Centro Comunal Zonal 6 de Montevideo)

Federico Alvarez - Fernando Massa

Tutores:
Ramón Alvarez
Laura Nalbarte
Licenciatura en Estadística

Puntaje .....
Tribunal
Profesor.....(nombre y firma).
Profesor.....(nombre y firma).
Profesor.....(nombre y firma).
Fecha.....

Agradecimientos

- A nuestras familias, ya que sin su constante apoyo, esta tarea habría sido irrealizable.
- A las compañeras del Instituto de Higiene y la Escuela de Nutrición quienes siempre tuvieron la mejor disposición para entender nuestras dudas y por proporcionarnos los elementos y los conceptos necesarios para una mejor comprensión del tema.
- A nuestros tutores: Laura Nalbarte y Ramon Alvarez por su dedicación y sugerencias realizadas.
- A nuestros compañeros de trabajo y amigos por el apoyo incondicional.

Resumen

En este estudio se analizó en que grado afectan los parásitos intestinales al estado nutricional y antropométrico de los niños de la escuela 317 (Zonal 6 de Montevideo). La investigación fue diseñada para llevarse a cabo en dos momentos del tiempo con una intervención de por medio. Considerando la estructura de dependencia de los datos, las técnicas utilizadas fueron de carácter longitudinal. Las principales conclusiones obtenidas fueron que pese a que el período en el que los niños estuvieron libres de parásitos fue breve, este fue suficiente para que, en promedio, se diera un leve cambio significativo en las variables que resumen el estado antropométrico de esta población. También se observó que la tenencia o no de saneamiento afectó tanto la situación antropométrica y parasitaria inicial de los niños como el desarrollo antropométrico de los mismos entre etapas.

Palabras claves: Parásitos Intestinales, Datos Longitudinales, Estado Nutricional.

Índice general

Índice de figuras

Índice de cuadros

Parte I

Introducción

Capítulo 1

Introducción

Las enteroparasitosis (infecciones que se localizan en el intestino) pueden transcurrir durante largo tiempo asintomáticas sin diagnosticar. Pero también pueden llegar a provocar cuadros digestivos, inclusive con severa repercusión sobre el crecimiento y desarrollo en los niños. Actualmente se está investigando la incidencia que pueden tener las infecciones parasitarias intestinales sobre el rendimiento escolar, por ejemplo a través de la irritabilidad y el cansancio que provocan, con repercusión sobre la capacidad intelectual y la atención. El comportamiento humano tiene gran importancia en la transmisión de las infecciones intestinales por parásitos, por lo tanto el éxito de las medidas de control que se implementen dependerá en gran medida de la modificación que se obtenga de los hábitos de comportamiento humano en el sentido de promover la salud y no contribuir a deteriorarla. Los objetivos de los programas de control de las enteroparasitosis están dirigidos fundamentalmente a reducir la morbilidad a corto plazo mediante la atención médica individual, quimioterapia y educación sanitaria. A largo plazo pretenden reducir la prevalencia a través de la mejora del saneamiento, abastecimiento de agua potable y promoción de la higiene personal y alimentaria.

En el documento presentado tras la conferencia de "Desparasitación para la Salud y el Desarrollo" llevada a cabo en Ginebra entre el 29 y el 30 de noviembre del año 2004, [20], la sociedad para el control de parásitos (Partners for Parasite Control) realizó un conjunto de medidas a tomar en el ámbito gubernamental, de instituciones de salud y de investigación. En este último, se afirma la necesidad urgente de realizar investigaciones en pro del desarrollo de vacunas y tratamientos en contra de las infecciones parasitarias.

La investigación descrita en este documento surgió por parte de un proyec-

to financiado a través de la Comisión Sectorial de Investigación Científica (CSIC) en el llamado *Proyectos de Investigación e Innovación Orientados a la Inclusión Social*.

En relación con las parasitosis intestinales el problema que venía siendo detectado desde el año 2006, cuando maestras comunitarias concurrieron al Instituto de Higiene (Dpto.de Parasitología) a manifestar su preocupación porque los niños expulsaban gusanos. El Censo de Talla de escolares demostró un elevado porcentaje de niños de esa escuela con retraso en la talla.

Parte de la preocupación del equipo del Instituto de Higiene pasó por su creencia de que los padres no comprendían la repercusión que podía tener para sus hijos el parasitismo intestinal y no jerarquizaban en este contexto la problemática planteada ni la vinculaban con el crecimiento ni la capacidad de aprendizaje. Los vínculos con los escolares en muchas ocasiones resultaban desgastantes debido al elevado nivel de violencia tanto verbal como físicamente expresado entre ellos.

El proyecto consistió de un estudio longitudinal descriptivo donde se hicieron dos tomas de datos con una intervención entre ambas para medicar a los niños de acuerdo a la/s infección/es hallada/s. Acorde a Rodríguez y Llorca [22], desde un punto de vista epidemiológico un estudio longitudinal es sinónimo de estudio de cohortes o seguimiento, mientras que desde el punto de vista estadístico este tipo de estudio implica mediciones repetidas.

La estrategia planteada fue la siguiente. Se creó un formulario que relevase aspectos antropométricos, ambientales, sanitarios y algunas características enfocadas a los antecedentes de ciertas patologías. Adicionalmente, con el fin de conocer el estado sanitario de la población, se planteó la aplicación de exámenes coproparasitarios y espátulas adhesivas. Se entiende por examen coproparasitario aquellas técnicas diagnósticas que constituyen la indicación metodológica para la identificación de la mayoría de las enteroparasitosis motivadas por protozoarios o helmintos. La espátula adhesiva es el método de elección para el diagnóstico de oxiuros, permitiendo recoger e identificar los huevos puestos en el margen anal del paciente.

A partir de estas herramientas, se realizó en una primera instancia un sondeo general que permitiese establecer una descripción inicial del estado de los niños.

Una vez obtenidos los resultados referentes a las infecciones, se procedió a medicar a los niños que presentaron uno o más tipos de parásitos intestinales con la finalidad de evaluar, luego de seis meses, su evolución antropométrica.

Luego de dicho período se procedió a recoger los datos antropométricos y coproparasitarios para su posterior análisis individual y en conjunto con los recabados en la primer toma de datos.

1.1. Objetivos

El objetivo general del presente trabajo consistió en la evaluación del impacto generado por infecciones parasitarias intestinales en el estado nutricional de niños que concurren a la escuela 317 (CCZ 6 de Montevideo) durante el año 2009.

Para ello se plantearon los siguientes objetivos específicos:

- Conocer el estado antropométrico de la población bajo estudio. Bajo este apartado, las mediciones realizadas fueron transformadas en "Z scores" según la metodología propuesta por la Organización Mundial de la Salud (OMS).
- Conformar tipologías de individuos mediante técnicas de análisis multivariante tanto en lo que respecta al estado antropométrico como a la presencia de los distintos agentes parasitarios.
- Cuantificar la relación entre el estado antropométrico y la presencia de agentes parasitarios.
- Evaluar el estado nutricional luego del tratamiento antiparasitario correspondiente. Para ello se utilizaron distintos enfoques de análisis de datos longitudinales.

El documento se organiza de la siguiente manera: En el capítulo dos se exponen los principales antecedentes (en la región) sobre estudios de índole nutricional, así como también los fundamentos teóricos de esta investigación. En el capítulo tres se plantean someramente los procedimientos estadísticos utilizados para responder a los objetivos ya expuestos, dentro de dicho captulo los procedimientos principales se agrupan en test de medias multivariadas, test de homogeneidad marginal, anlálisis de cluster y análisis multivariado de la varianza. En el capítulo cuatro se exhiben tanto la estrategia de análisis como los resultados obtenidos en cada una de las etapas de la investigación. Finalmente, en el capítulo cinco se comentan las conclusiones obtenidas y las posibles líneas de investigación a futuro.

Parte II

Aspectos metodológicos

Capítulo 2

Antecedentes y Marco teórico

2.1. Antecedentes

Las infecciones parasitarias intestinales tienen tasas de prevalencia elevadas en varias regiones. En general tienen asociadas baja mortalidad, pero igualmente ocasionan importantes problemas sanitarios y sociales debido a su sintomatología y complicaciones. A continuación se presentan investigaciones de características similares a las del presente estudio.

En el año 2009, en la ciudad de La Plata, Argentina, Gamboa et al [12] investigan la relación entre las condiciones socio-ambientales y las infecciones parasitarias intestinales de niños menores a 12 años pertenecientes a un barrio carenciado en dicha ciudad. Valiéndose de herramientas de análisis factorial, se llegó a la conclusión de que las variables socio-ambientales presentaron mayor asociación con el retraso en el crecimiento y las infecciones provocadas por geohelmintos, sobre todo en los niños más carenciados. Adicionalmente el estudio concluyó que las condiciones en las que viven estos niños y familiares son muy variables y que esta variablidad puede atribuirse a la forma en que los padres utilizan sus escasos recursos para influir en la morbilidad de sus hijos.

Otro estudio similar fue realizado de manera conjunta entre la Intendencia Municipal de Montevideo (IMM) y la Facultad de Medicina en el año 1997. El mismo consistió de un plan que comprendió la realización de actividades informativas y educativas así como actividades de diagnóstico mediante exámenes coproparasitarios y espátulas adhesivas en dieciséis guarderías ubicadas en diferentes puntos de la periferia de la ciudad de Montevideo. Una

de las conclusiones resaltadas por Acuña et al [2][1] en esta investigación fue que un 45,5 % de los niños estudiados en la primera instancia presentó algún tipo de infección parasitaria. Por lo que se considera pertinente mantener informada a la población sobre este tipo de infecciones.

2.2. Parasitosis intestinal

Los helmintos o gusanos que parasitan el intestino humano, son importantes agentes de morbilidad y causa de mortalidad en amplias poblaciones de diversas regiones del planeta.

El impacto de los helmintos intestinales sobre la salud de la población suele quedar enmascarado por las dificultades de su diagnóstico, entre ellas las principales son:

- Inespecificidad de los síntomas.
- Carencia de laboratorios adecuados.
- Bajas cargas parasitarias sin expresión clínica.
- Dificultades para la consulta médica por parte del afectado.

Debe destacarse que estos parásitos presentan una mayor tasa de prevalencia sobre poblaciones con condiciones desfavorables tanto sanitarias como culturales y ambientales.

En cuanto a las condiciones generales se refiere, cabe destacar que estos parásitos tienen como principal factor común la necesidad de un alto grado de "fecalismo ambiental". Es decir, que por carencias de saneamiento y deficiente abastecimiento de agua potable, el ambiente, agua y alimentos tienen un elevado índice de contaminación con excretas humanas facilitando así la transmisión.

Las helmintiasis intestinales presentan (sobre todo las geohelmintiasis) un aumento preocupante en su frecuencia vinculado a la situación de riesgo social y deterioro sanitario que viven algunos sectores de la sociedad. Las infecciones intestinales por helmintos que se observan en el Uruguay son producidas fundamentalmente por los siguientes agentes:

- *Enterobios vermicularis* (oxiuro)
- *Ascaris lumbricoides*

- *Trichuris trichuria* (tricocéfalo)
- *Strongyloides stercoralis*
- *Hymenolepis nana*
- *Taenia saginata* (solitaria)

En cuanto a los mecanismos de transmisión pueden definirse tres grandes categorías:

- Geohelmintos. Son formas que habitan en el suelo y penetran en su huésped (persona que permite el alojamiento de un agente infeccioso) ya sea por vía oral o transcutánea.
- Helmintos de transmisión directa. Penetran en su huésped a través del ciclo fecal-oral o ano-mano-boca.
- Helmintos transmitidos por carnivorismo. Son formas infectantes que habitan en la carne vacuna mal cocida.

Los síntomas y signos habituales son, en general, inespecíficos y de difícil definición clínica. No obstante, estos parásitos pueden condicionar la vida de las personas afectando su estado nutricional y su desarrollo, alterando sus procesos cognitivos o provocando complicaciones riesgosas. Las manifestaciones clínicas son de diversa índole, las más comunes son alérgicas (urticarias), neurológicas (cefaleas), generales (disminución de peso), hematológicas (anemia) y digestivas (mal absorción de nutrientes). Los métodos paraclínicos de elección para el diagnóstico de las helmintiasis intestinales (que de describen brevemente a continuación) son el examen coproparasiario y la espátula adhesiva.

- exámen coproparasitario. El exámen coporparasitario comprende un conjunto de métodos para la observación macroscópica y microscópica de las heces, incluyendo métodos de concentración de los elementos parasitados y de coloraciones específicas, los cuales permiten poner en evidencia a huevos, larvas y helmintos adultos.
- espátula adhesiva. La espátula adhesiva es el método de elección para el diagnóstico de oxiuros, permitiendo recoger e identificar los huevos puestos en el margen anal del paciente.

Ambos métodos paraclínicos son realizadas serialmente de manera de aumentar las posibilidades de diagnóstico.

Capítulo 3

Marco metodológico

En este capítulo se exponen las técnicas y procedimientos utilizados a lo largo del estudio para responder a los objetivos planteados. El siguiente compilado pretende esbozar una breve descripción de los métodos estadísticos utilizados y exponer la justificación de la elección de cada uno de ellos.

3.1. Estandarización

Existen tres sistemas mediante los cuales un niño o un grupo de niños pueden ser comparados a una población de referencia: los percentiles, el porcentaje a la mediana y los puntajes Z (Z scores). Tanto en encuestas poblacionales como de vigilancia epidemiológica, el sistema de los Z scores es ampliamente reconocido como el mejor para el análisis de datos antropométricos por sus ventajas ante los otros métodos.

En este sistema, la talla, el peso y el índice de masa corporal (IMC) para la edad se interptetan en términos del número de desvíos standard por exceso o defecto en torno a una media. La interpretación de resultados en términos de Z scores tiene varias ventajas:

- La escala de los mismos es lineal, por ende un intervalo fijo de Z scores corresponde a un intervalo fijo en la talla medida en cm, o en el peso medido en kg, o en el IMC medido en $\frac{kg}{m^2}$ para todos los niños de la misma edad (medida en meses). En otras palabras, los Z scores presentan la misma relación estadística en relación a la distribución de referencia en torno a la media en todas las edades, lo cual hace que los resultados sean comparables a lo largo de todas las edades.

- Son independintes del sexo, es decir se calculan de forma separada para niños y niñas permitiendo la comparación entre ellos.

La OMS, en colaboración con instituciones de varios países, realizó un estudio multi-céntrico para elaborar nuevas referencias del crecimiento para lactantes y niños pequeños, el mismo se conoce como el Estudio Multi-Centro de las Referencias del Crecimiento (MGRS) de la OMS. El MGRS que abarcó 8500 niños, se llevó a cabo en seis países que representan las principales regiones del mundo. Para ello se utilizó el modelo LMS desarrollado por Cole [8], el cual fué una herramienta capaz de proporcionar medidas estándar a través de tres parámetros (potencia, mediana y coeficiente de variación). Uno de los resultados de dicho estudio fué un conjunto de parámetros para cada edad (medida en meses) para los siguientes indicadores antropométricos: talla, peso e IMC.

El Z score puntaje z para la medición y, se calcula de la siguiente forma:

$$\frac{[y/M(t)]^{L(t)} - 1}{S(t)L(t)} \tag{3.1}$$

Donde $L(t)$,$M(t)$ y $S(t)$ son los parámetros ya mencionados de potencia, mediana y coeficiente de variación respectivamente para la edad t. En el apéndice A.1 pueden encontrarse algunos valores de $L(t)$,$M(t)$ y $S(t)$ y el código utilizado en R para calcularlos.

En la presente investigación el uso de estos indicadores permitió comparar el estado nutricional de uno o varios niños en distintos momentos del tiempo. No obstante la distribución de estos indicadores se aleja sensiblemente de la normalidad en las colas de la distribución. Es por esto que se propuso el siguiente ajuste.

$$\mathrm{z}^* = \begin{cases} z & si \quad |z| < 3 \\ 3 + (\frac{y - SD3pos}{SD23pos}) & si \quad z > 3 \\ -3 + (\frac{y - SD3neg}{SD23neg}) & si \quad z < -3 \end{cases} \tag{3.2}$$

Donde $SD3pos$, $SD3neg$, $SD23pos$ y $SD23neg$ se detallan a continuación:

$$\begin{aligned} &SD3pos\text{=M(t)[1+3L(t)S(t)]}^{1/L(t)} \\ &SD3neg\text{=M(t)[1-3L(t)S(t)]}^{1/L(t)} \\ &SD23pos\text{=M(t)[1+3L(t)S(t)]}^{1/L(t)} - M(t)[1+2L(t)S(t)]^{1/L(t)} \\ &SD23neg\text{=M(t)[1-2L(t)S(t)]}^{1/L(t)} - M(t)[1-3L(t)S(t)]^{1/L(t)} \end{aligned} \tag{3.3}$$

De esta forma se logra que los puntajes Z se asemejen más a una distribución normal.

3.2. Test de medias multivariado

El estado antropométrico de los niños (peso, talla e IMC) se caracterizó a través de tres Z scores, el test de medias multivariado se utilizó para evaluar el peso, talla e IMC de los niños y las posibles diferencias que podrían surgir debido a variables como la tenencia de saneamiento, teniendo en cuenta la marcada estructura de correlación entre las variables.

En esta sección se desarrolla el caso en que se desea poner a prueba una hipótesis sobre las medias de un conjunto de p variables de forma conjunta. A continuación se detallan cuatro motivos para tener en cuenta tests multivariados:

- Al usar p pruebas univariadas, el error de tipo I aumenta al aumentar el número de variables. Por otro lado, los test multivariados preservan el nivel α.
- Los tests univariados no toman en cuenta las correlaciones existentes entre las variables.
- En muchos casos, los tests multivariados son más potentes que los univariados, es más, suele suceder que las p pruebas univariadas no logran ser significativas, mientras que en las pruebas multivariadas se combinan pequeños efectos de cada una de las variables para lograr significación.
- Muchos tests multivariados sobre las medias tienen el extra de que a partir de los mismos se puede concluir cual/es de las variables influyen en mayor medida sobre la significación de la prueba.

La figura 3.1 ilustra el tercer punto. En el mismo se superponen las regiones de aceptación para un nivel de 95 % de un test bivariado (elipse) y dos test univariados (rectángulo) para las variables y_A y y_B

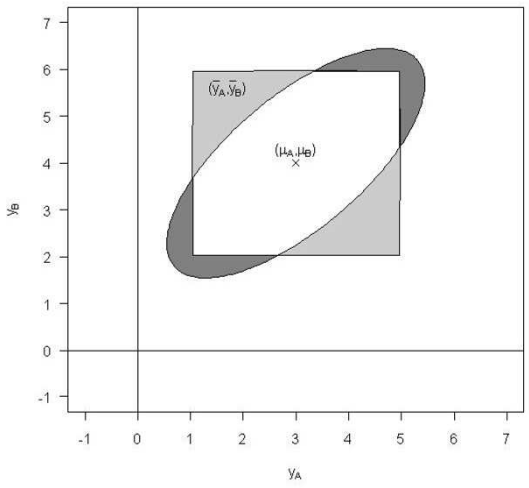

Figura 3.1: Región de rechazo multivariada

3.2.1. Estadístico T^2 para dos muestras

Se considerara el caso en el que se dispone de p variables para cada una de las n unidades de muestreo correspondientes a dos grupos. En este caso, interesa probar la siguiente hipótesis:

$$H_0)\ \mu_A = \mu_B$$

contra

$$H_1)\ \mu_A \neq \mu_B$$

donde, tanto μ_A como μ_B, son vectores de medias poblacionales de tamaño p. Así $y_{A1}, y_{A2}, \ldots, y_{An_A}$ es una muestra aleatoria correspondiente a una distribución normal p-variada de media μ_A y matriz de covarianzas Σ_A. Por otro lado, $y_{B1}, y_{B2}, \ldots, y_{Bn_B}$ es una muestra aleatoria correspondiente a otra distribución normal p-variada de media μ_B y matriz de covarianzas Σ_B.

Se supone la independencia entre ambas muestras y la igualdad de las matrices de covarianzas. El cumplimiento de estos supuestos es necesario para que la distribución del estadístico T^2 sea $T^2_{g1,g2}$ de Hotelling, siendo g1 y g2 los grados de libertad de la distribución.
Así el estadístico de prueba adopta la siguiente forma:

$$T^2 = \frac{n_A n_B}{n_A + n_B}(\bar{y}_B - \bar{y}_A)' \; S^{-1}_{pool}(\bar{y}_B - \bar{y}_A)^1 \tag{3.5}$$

Donde $\bar{y}_A$ y $\bar{y}_B$ son los vectores de medias muestrales de cada uno de los grupos. S_{pool} es un estimador insesgado de la matriz de covarianzas común a ambos grupos.

$$S_{pool} = \frac{(n_A - 1)S_A + (n_B - 1)S_B}{n_A + n_B - 2} \tag{3.6}$$

S_A y S_B son las matrices de covarianzas muestrales de los grupos A y B respectivamente.

Bajo el cumplimiento de la hipótesis nula, el estadístico se distribuye T^2_{p,n_A+n_B-2}.

Donde $(1/n_A + 1/n_B)S_{pool}$ es la estimación de la matriz de covarianzas muestral de $\bar{y}_A - \bar{y}_B$ y S_{pool} es independiente de $\bar{y}_A - \bar{y}_B$ debido a que las observaciones son normales multivariadas.

[1] Algunas características de este estadístico se detallan a continuación:

- Para que S_{pool} sea no singular es necesario que $n_A + n_B - 2 > p$
- El estadístico T^2 es un escalar.
- Incluso cuando la hipótesis alternativa de la prueba es bilateral, la región crítica es unilateral.
- Para simplificar el cálculo de p-valores, el estadístico T^2 puede ser transformado a un estadístico F de la siguiente manera:

$$\frac{n_A + n_B - p - 1}{(n_A + n_B - 2)p} T^2 \sim F_{p,n_A+n_B-2} \tag{3.5}$$

3.2.2. Procedimientos post hoc

En el caso en que se rechace la hipótesis de igualdad de medias, los procedimientos que se describen a continuación pueden ser útiles para detectar cual o cuales de las variables involucradas determinan en mayor medida este hecho. En este estudio en particular, la elección de estos procedimientos responde a la necesidad de explorar cual/es de los puntajes Z es/son el/los más influyente/s tras el rechazo de H_0. Algunos de los procedimientos son:

- Pruebas t univariadas para cada variable. Estas pruebas suelen llamarse "*protegidas*" debido a la previa significación del estadístico T^2. En Hummel y Sligo [14] se recomienda no ajustar el nivel de cada prueba ya que sus experimentos demuestran que (para un nivel de 5 % en cada prueba) se logra un nivel global muy cercano al nivel nominal de 5 %.
- Pruebas t o F parciales. En estos se lleva acabo la prueba en cada variable ajustada por el efecto de las demás.
- Análisis de los coeficientes de la función discriminante (ver anexo B.4). Al examinar el valor absoluto de dichos coeficientes, se puede establecer cual/es de las variable/s influye de mayor o menor manera sobre el rechazo de H_0

3.2.3. Pruebas para observaciones apareadas

Al ser este un estudio constituido por dos etapas, se dispuso de mediciones antropológicas en dos momentos para cada individuo, parece natural pensar en las llamadas "*pruebas para datos relacionados*"para evaluar el posible efecto de la intervención en cambio de las variables.

El siguiente procedimiento es de especial utilidad en los casos en que la i-ésima observación del primer grupo no sea independiente de la i-ésima observación del segundo grupo. Un ejemplo clásico de esta situación se da cuando dos tratamientos son aplicados al mismo individuo. En virtud de este "*apareamiento*" los dos grupos de observaciones se encuentran correlacionados y por ende las pruebas descritas anteriormente no son apropiadas. La solución a este problema es trabajar con las diferencias entre las observaciones apareadas de los grupos. Adicionalmente es necesario plantear el supuesto de que la diferencia entre los grupos de observaciones tienen una distribución normal p-variada, es decir:

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \sim N_{2p}\left(\begin{pmatrix} \mu_A \\ \mu_B \end{pmatrix}, \begin{pmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{pmatrix}\right) \tag{3.7}$$

De esta manera, en vez de poner a prueba la hipótesis $H_0)\ \mu_A = \mu_B$ se testea la siguiente hipótesis nula $H_0)\ \mu_d = 0$, siendo $d = y_B - y_A$ el vector de las diferencias de las variables entre los grupos. Un aspecto interesante sobre esta prueba es que no necesita que las matrices de covarianzas de los grupos sean iguales, es decir Σ_{AA} puede ser distinta de Σ_{BB}.
Así, pese a que en un principio la prueba involucra dos grupos, el estadístico a utilizar es el T^2 para un solo grupo, siendo este el de las diferencias.

Vale señalar que, en caso de que la prueba resulte ser significativa, los procedimientos señalados en el apartado anterior siguen siendo válidos para los tipos de pruebas considerados en este apartado.

3.3. Test de homogeneidad marginal

Así como el estadístico de Hotelling pretende detectar el cambio producido sobre las variables cuantitativas (Z scores), las pruebas descritas en esta sección del estudio son utilizadas para evaular el cambio sobre variables categóricas. Esta prueba se usó para el caso concreto de la presencia/ausencia de cada uno de los diferentes parásitos en los indivuduos estudiados antes y después de la intervención.

En el caso de que se disponga de una variable de respuesta categórica (con I clases) para un cierto número de individuos en dos momentos del tiempo, la técnica a utilizar es llamada *matched paired analysis*. Debido al apareamiento de los individuos en las muestras (momentos), los datos son estadísticamente dependientes. El procedimiento a seguir para analizar este tipo de datos es construir una tabla de contingencia con las mismas categorías en filas y columnas, esta tabla suele llamarse *tabla cuadrada*.

		Y_1			
		n_{11}	n_{12}	n_{13}	$n_{1.}$
Y_2		n_{21}	n_{22}	n_{23}	$n_{2.}$
		n_{31}	n_{32}	n_{33}	$n_{3.}$
		$n_{,1}$	$n_{,2}$	$n_{,3}$	$n_{..}$

Esta tabla corresponde a la situación en la que una variable con tres categorías ($I = 3$) se registra sobre $n_{..}$ individuos en dos momentos. De esta forma, n_{ij} representa la cantidad de observaciones que pertenece a la categoría i en el primer momento y j en el segundo. Por otro lado, $n_{i.}$ es la cantidad de individuos que adoptaron el valor i en el primer momento y $n_{.j}$ es la cantidad que respondieron j en el segundo. Al dividir los elementos de la tabla entre $n_{..}$ se obtienen los valores π_{ij}. Los mismos estiman la probabilidad de que un individuo adopte el valor i en el primer momento y j en el segundo.

La homogeneidad marginal refiere a la falta de diferencia significativa entre una o más de las proporciones marginales de las filas con su homóloga en las columnas. Estrictamente hablando, esta situación corresponde a que:

$$\pi_{i.} = \pi_{.i} \quad \forall\ i = 1, 2, \ldots, I \tag{3.8}$$

siendo esta la hipótesis nula de la prueba.

En cuanto al estadístico de prueba, existen dos alternativas. La clave del mismo reside en el vector d. Este vector recoge las diferencias entre las probabilidades marginales, es decir:

$$d_i = \pi_{i.} - \pi_{.i} \tag{3.9}$$

Vale aclarar que la dimensión del mismo es $I - 1$. El I-ésimo elemento es redundante debido a que $\sum_{i=1}^{I} d_i = 0$. Bajo la hipótesis de homogeneidad marginal, $E(d) = 0$. No obstante, es la matriz de covarianzas V la que origina las dos alternativas. Una de las alternativas, propuestas por Stuart [25] y Maxwell [16] es la siguiente matriz de covarianzas:

$$V_{ij}^{SM} = \begin{cases} -(\pi_{ij} + \pi_{ji}) & \text{cuando } i \neq j \\ \pi_{i.} + \pi_{.i} - 2\pi_{ii} & \text{cuando } i = j \end{cases} \tag{3.10}$$

La matriz de covarianzas propuesta por Bhapkar [5] es la siguiente:

$$V_{ij}^{B} = \begin{cases} -(\pi_{ij} + \pi_{ji}) - (\pi_{.i} + \pi_{i.})(\pi_{.j} + \pi_{j.}) & \text{cuando } i \neq j \\ \pi_{i.} + \pi_{.i} - 2\pi_{ii} - (\pi_{.i} + \pi_{i.})^2 & \text{cuando } i = j \end{cases} \tag{3.11}$$

Luego, dada la normalidad multivariada asintótica del vector d, el siguiente estadístico posee una distribución aproximadamente chi-cuadrada con $I-1$ grados de libertad.

$$W_0 = n_{..} d' V^{-1} d \sim \chi^2_{I-1} \tag{3.12}$$

Pese a que el estadístico no proporciona el mismo valor según la forma que adopte la matriz de covarianzas utlizada, la elección de dicha matriz proporciona resultados asintóticamente equivalentes. No obstante, la alternativa proporcionada por Bhapkar es más potente y por lo tanto suele ser preferida.

3.4. Análisis de cluster

Con el propósito de cumplir con el objetivo en el cual se planteaba construir tipologías referentes al estado parasitario y antropométrico de la población se utilizaron técnicas de análisis de cluster. Dada la naturaleza cuantitativa de los Z scores, los métodos mediante los cuales se agruparon los individuos fueron los jerárquicos agregativos, mas concretamente el método de Ward. Por otro lado, como las variables parasitarias eran binarias (si/no) el método utlizado para conformar los grupos fué el de cluster basado en un modelo probabilístco.

Según se describe en Blanco [6], el análisis de cluster, también conocido como análisis de conglomerados, es una técnica estadística multivariante cuya finalidad es dividir un conjunto de objetos en grupos de manera que los perfiles de los objetos en un mismo grupo sean muy similares entre sí y los de los objetos de clusters diferentes sean lo mas distintos posible. En el anexo B.1 se encuentra desarrollada esta sección con mayor profundidad.

3.5. Análisis de cluster probabilístico

Siguiendo la metodología propuesta en Moustaki y Papageorgiou [18], en los modelos de clase latente se asume la existencia de una variable aleatoria Z que consta de G clases, tal que:

$$P(Z = j) = \tau_j,\ \sum_{j=1}^{G} \tau_j = 1 \tag{3.13}$$

Se denotará por x_{ik} al i-ésimo elemento de la k-ésima variable, siendo $(x_{i1}, x_{i2}, \ldots, x_{ip})$ el patrón de respuesta del i-ésimo individuo.

Sea $(x_1, x_2, \ldots, x_p)$ un vector de p variables explicativas con distribución Bernoulli con parámetro ϕ_{jk} dónde el subíndice k hace referencia a cada una de las variables y j a cada uno de los grupos que se introducirán a continuación.

Las probabilidades τ_j suelen llamarse probabilidades a priori. La distribución conjunta de las variables observadas está dada por la siguiente mezcla finita probabilística:

$$f\left(x_i\right) = \prod_{j=1}^{G} \prod_{k=1}^{p} [\tau_j \phi_{jk}^{x_{ik}} (1 - \phi_{jk})^{(1-x_{ik})}]^{I(Z_i=j)} \tag{3.14}$$

Es necesario incluir la variable aleatoria $I(Z_i = j)$ debido a que no se sabe a que grupo pertenece cada observación. Finalmente la log-verosimilitud de una muestra de n individuos está dada por la siguiente expresión:

$$L(\theta|x) = \sum_{i=1}^{n} \sum_{j=1}^{G} I_{(Z_i=j)} \sum_{k=1}^{p} [\log(\tau_j) + x_{ik} \log(\phi_{jk}) + (1 - x_{ik}) \log(1 - \phi_{jk})] \tag{3.15}$$

Donde el vector θ contiene todos los parámetros del modelo. Los parámetros se estiman iterativamente a través del algoritmo EM [11]. Para este caso, cada iteración de dicho algoritmo, consta de las siguientes etapas:

1. paso E) $Q(\theta|\theta^t) = E_{Z|X,\theta^t}[L(\theta|x)]$
2. paso M) $\theta^{t+1} = \underset{\theta}{\max}\, Q(\theta|\theta^t)$

Este proceso iterativo se repite hasta que el cambio en la log-verosimilitud sea menor que cierta tolerancia pre-especificada. Cabe señalar que uno de los principales inconvenientes de este algoritmo es su dependencia del valor

inicial del vector de parámetros, para hacer frente a este problema en este trabajo se optó por la estrategia de inicializar el algoritmo con distintas valores iniciales y optar por los resultados del modelo que obtenga el mayor valor de la log-verosimilitud.

Algunas de las ventajas de este enfoque son las siguientes:

- Dada la naturaleza paramétrica del modelo, se puede seleccionar el número de grupos utilizando criterios de información como el Bayesian Information Criterion (BIC) o el Akaike Information Criterion (AIC) (véase apéndice B.3).
- A través de la distribución a posteriori de la variable aleatoria Z condicional a los parámetros y a las variables binarias, se puede definir el grado de pertenencia de cada individuo a cada grupo.

$$P(Z_i = j) = \frac{\tau_j \prod_{j=1}^{G} \phi_{jk}^{x_{ik}} (1 - \phi_{jk})^{1-x_{ik}}}{\sum_{j=1}^{G} \tau_j \prod_{j=1}^{G} \phi_{jk}^{x_{ik}} (1 - \phi_{jk})^{1-x_{ik}}} \tag{3.16}$$

- Mediante los parámetros del modelo se pueden calcular frecuencias esperadas que, comparadas con las frecuencias observadas, indiquen el grado de ajuste del modelo.

3.6. Análisis multivariado de la varianza (MANOVA)

El principal objetivo del estudio es analizar la asociación entre los estados parasitario y antropológico de los niños. A partir de los grupos parasitarios generados por la técnica de análisis de cluster probabilístico, se decidió investigar la diferencia entre los vectores de medias de las variables antropométricas.

El análisis multivariado de la varianza es el equivalente multivariado de la comparación de varios grupos realizada por el ANOVA. El propósito final del MANOVA es testear si los vectores de medias de dos o más grupos provienen o no de la misma distribución.

El modelo para p variables es el siguiente:

$$y_{ijr} = \mu_r + \alpha_{ir} + \epsilon_{ijr} \tag{3.17}$$
$$i = 1, 2, \ldots, k \quad j = 1, 2, \ldots, n_i \quad r = 1, 2, \ldots, p$$

Interesa investigar si existen diferencias significativas entre los k vectores de medias. La hipótesis nula es:

$$H_0)\ \alpha_{1r} = \alpha_{2r} = \ldots = \alpha_{kr} \tag{3.18}$$

El cumplimiento de la hipótesis nula requiere que se cumplan las $p(k-1)$ igualdades, mientras que el incumpliendo de una sola desigualdad implicará la falsedad de la hipótesis.

De manera análoga a la suma de cuadrados "entre" y "dentro" en el ANOVA, las siguientes matrices contienen sumas de cuadrados y productos "entre" y "dentro":

$$H = \sum_{i=1}^{k} n_i \left(y_{i.} - y_{..}\right)\left(y_{i.} - y_{..}\right)' = \sum_{i=1}^{k} \frac{1}{n_i} y_{i.} y_{i.}' - \frac{1}{\sum_{i=1}^{k} n_i} y_{..} y_{..}' \tag{3.19}$$

$$E = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_{i.}\right)\left(y_{ij} - \bar{y}_{i.}\right)' = \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij} y_{ij}' - \sum_{i=1}^{k} \frac{1}{n_i} y_{i.} y_{i.}' \tag{3.20}$$

La matriz de hipótesis H, contiene en su diagonal las sumas de cuadrados "entre" de cada una de las p variables. Los elementos fuera de la diagonal principal son las sumas de productos de cada par de variables. Asumiendo que no existe dependencia lineal entre las variables, el rango de la matriz H se deduce como mínimo entre p y $k-1$. Es por esto que la matriz H es singular. Por otro lado la matriz de error E contiene en su diagonal la suma de cuadrados "dentro" de cada una de las p variables y fuera de la misma se encuentran las sumas de productos. El rango de E es p, salvo que los grados de libertad del error sean menos que p.

En base a estas matrices se definen los siguientes estadísticos:

1. Estadístico de Wilks:

   Definición:

   El estadístico Λ de Wilks, que pone a prueba la hipótesis de igualdad de vectores de medias, es el siguiente:

$$\Lambda = \frac{|E|}{|E+H|} \tag{3.21}$$

   Una particularidad de esta prueba es que la hipótesis nula se rechaza cuando Λ adopta un valor menor a un valor crítico determinado por p, gl_H y gl_E. Los valores críticos para el test se pueden encontrar en Wall [27]. La forma de proceder de este estadístico consiste en comparar la variablidad "entre" con la variabilidad total.

   Características:

   - Λ puede ser expresado en términos de los valores propios de la matriz $E^{-1}H$ de la siguiente manera:

$$\Lambda = \prod_{i=1}^{p} \frac{1}{1+\lambda_i} \tag{3.22}$$

   - El rango en el cual varía Λ es el intervalo $[0, 1]$ y el test basado en este estadístico se llama "inverso", en el sentido de que se rechaza la hipótesis nula para valores cercanos a 0. Esto sucede cuando los vectores de medias muestrales están lo suficientemente dispersos comparados con la variación dentro de los grupos. Esto sucede cuando H constituye una gran parte de la variación total haciendo que Λ se aproxime a 0.

2. Estadístico de Roy:

   Definición:

   Esta prueba trabaja sobre una combinación lineal de las p variables dependientes. Una vez determinado el vector a, se obtienen las medias "transformadas" $\bar{z}_{i.} = a^{'}\bar{y}_{i.}$ tal que este maximise la dispersión de las

mismas. De esta manera, a es el vector propio asociado al primer valor propio de la matriz $E^{-1}H$. El estadístico basado en dicho valor propio es el siguiente:

$$\theta = \frac{\lambda_1}{1 + \lambda_1} \tag{3.23}$$

Los valores críticos para el test se pueden encontrar en Parson y Hartley [19]. La distribución de este estadístico está indexada por tres parámetros, s, m y N.

$$s = min\left(gl_H, p\right) \tag{3.24}$$

$$m = \frac{1}{2}\left(|gl_H - p| - 1\right) \tag{3.25}$$

$$N = \frac{1}{2}\left(gl_H - p - 1\right) \tag{3.26}$$

Características:

- Los coeficientes del vector a pueden ser inspeccionados para determinar cuales son las variables que contribuyen mayormente a la separación de las medias

3. Estadístico de Lawley-Hotelling:

   Definición:

   El estadístico U de Lawley-Hotelling se define como:

$$U = tr\left(E^{-1}H\right) \tag{3.27}$$

Este también es conocido como estadístico T^2 generalizado de Hotelling debido a que en el caso de que sólo se cuente con dos grupos el estadístico se reduce al T^2. Los valores críticos para el test se pueden encontrar en Davis [26].

Características:

- El estadístico U puede ser expresado en función de los p valores propios de $E^{-1}H$, esto es:

$$U = \sum_{i=1}^{p} \lambda_i \tag{3.28}$$

4. Estadístico de Pillai:

Definición:

Algunos investigadores consideran a esta prueba como la más potente y robusta de las cuatro. El estadístico V de Pillai es el siguiente:

$$V = tr\left[(E+H)^{-1} H\right] \tag{3.29}$$

La hipótesis nula es rechazada cuando V toma valores mayores a los valores críticos indicados en Schuurmann et al [23]. Al igual que en el estadístico de Roy, la distribución de este estadístico está indexada por tres parámetros, s, m y N.

El estadístico de Pillai puede ser considerado una extensión del de Roy. Esto se debe a que Roy considera sólo al máximo valor propio, mientras que Pillai considera relevante la información contenida en todos los valores propios de la matriz $E^{-1}H$.

Características:

- El estadístico V puede ser expresado en función de los p valores propios de $E^{-1}H$, esto es:

$$V = \sum_{i=1}^{p} \frac{\lambda_i}{1+\lambda_i} \tag{3.30}$$

3.6.1. Comparación de los estadísticos

En el caso de $p = 1$ (ANOVA) las medias poblacionales pueden ser ordenadas en una dimensión y por ende, solo sería necesaria una dirección para establecer si existen diferencias entre los grupos. En el caso multivariado, los vectores

de medias pertenecen a un subespacio de dimensión $s = min\,(p, k-1)$, es por esto que, dependiendo de la configuración de los mismos en este subespacio, podrán necesitarse (o no) más direcciones para "separar" los grupos.

Una manera de investigar la forma en que los vectores de medias están dispersos, es a través de los valores propios de $E^{-1}H$. Es así que, si solamente hay un valor propio "grande" y los restantes son muy "cercanos" a cero, entonces los vectores de medias estarán muy cerca de ser colineales. Análogamente si se tiene dos valores propios "grandes" y los restantes "cercanos" a cero, entonces los vectores de medias estarán muy cerca de conformar un plano, y así sucesivamente. Debido a que el test de Roy sólo toma en cuenta el máximo autovalor, es el uniformemente más potente en el caso en que los vectores de medias son colineales. Por otro lado, cuando los vectores de medias están más "dispersos", los otros tres estadísticos son más potentes. En cuanto a la robustez de los mismos frente a la violación de los supuestos de homogeneidad de matrices de varianzas y covarianzas y multinormalidad, se puede afirmar que cuando las matrices $\Sigma_1, \Sigma_2, \ldots, \Sigma_k$ se alejan de la situación de homogeneidad, el error de tipo I se ve afectado de diferente manera en cada uno de los estadísticos. Así, el estadístico menos afectado por el no cumplimiento de este supuesto es el de Pillai, seguido por el de Wilks, el de Hotelling-Lawley y en último lugar el de Roy. Por lo general, si los grupos están balanceados (en cuanto a su tamaño), los cuatro estadísticos son lo suficientemente robustos respecto a la homocedasticidad. En conclusión, el estadístico de Roy posee un buen desempeño en la situación de colinealidad de los vectores de medias y bajo el cumplimiento de los supuestos mencionados anteriormente. Para chequear el cumplimiento de este supuesto se puede utilizar la prueba de Box-M (Véase apéndiceB.2).

En cuanto a la falta de multinormalidad, las pruebas de Pillai, Hotellin-Lawley y Willks se desempeñan de mejor manera que la prueba de Roy. Para chequear el supuesto de multinormalidad, podrán utilizarse pruebas de normalidad multivariada como las de Mardia o Doornik-Hansen(véase apéndice A.2). En líneas generales los estadísticos que presentan mejores características son Pillai y Wilks.

3.6.2. Medidas multivariadas de asociación

Pese a que existen varias formas de cuantificar la asociación entre las variables dependientes y los diferentes factores, se expondrán sólo las dos más comunes.

- η^2_Λ generalizado. En 1932, Wilks propuso el siguiente estadístico:

$$\eta^2_\Lambda = 1 - \Lambda \tag{3.31}$$

Siendo Λ el estadístico de prueba de Wilks. Puede verse como, al aumentar la dispersión entre las medias, el valor del estadístico aumenta.

- η^2_θ de Roy

  El estadístico de Roy en sí mismo constituye una medida de asociación ya que puede demostrarse que equivale al cociente de las sumas de cuadrados "entre" y "dentro" de la variable $z = a_1^{'} y$.

$$\eta^2_\theta = \frac{\lambda_1}{1 + \lambda_1} \tag{3.32}$$

3.6.3. Manova mixto

Al tratarse de un estudio longitudinal, mediciones de un mismo individuo están correlacionadas en el tiempo. Es así que una de las posibles maneras de examinar el impacto de ciertas covariables en la evolución de los Z scores, es introducir un efecto aleatorio inducido por los individuos.

Ya sea para modelos de efectos fijos, aleatorios o mixtos, puede demostrase que existe un equivalente multivariado para el ANOVA. En el MANOVA mixto balanceado, las matrices de cuadrados medios esperados exhiben el mismo patrón que los cuadrados medios esperados del ANOVA. A continuación se detallan las características para un modelo mixto, tanto en el caso univariado como en el multivariado:

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk} \tag{3.33}$$

$$i = 1, 2, \ldots, a \quad j = 1, 2, \ldots, b \quad k = 1, 2, \ldots, n$$

$$\epsilon_{ijk} \sim N(0, \sigma^2)$$
$$\alpha_i \sim N(0, \sigma^2_\alpha)$$
$$(\alpha\beta)_{ij} \sim N\left(0, \tfrac{b}{b-1}\sigma^2_{\alpha\beta}\right)$$

$$Y_{ijk} = \mu + A_i + B_j + (AB)_{ij} + \epsilon_{ijk} \tag{3.34}$$

$$i = 1, 2, \ldots, a \quad j = 1, 2, \ldots, b \quad k = 1, 2, \ldots, n$$

$$\epsilon_{ijk} \sim N_p\left(0, \Sigma\right)$$
$$\mathrm{A}_i \sim N_p\left(0, \Sigma_A\right)$$
$$(AB)_{ij} \sim N_p\left(0, \tfrac{b}{b-1}\Sigma_{AB}\right)$$

En el caso univariado α es un factor aleatorio, β es fijo y la interacción constituye también un factor aleatorio, mientras que en el caso multivariado A es un factor aleatorio, B es un factor fijo y la interacción es un factor aleatorio. En ambos casos, α(A) es un factor aleatorio, β(B) es fijo y la interacción constituye también un factor aleatorio. A continuación, en el cuadro 3.1 se detalla el caso univariado incluyendo los cuadrados medios esperados para cada una de las fuentes de variación.

Fuente	gl	Cuadrados Medios Esperados
α	$a-1$	$\sigma^2 + nb\sigma^2_\alpha$
β	$b-1$	$\sigma^2 + na\frac{\sum_{j=1}^{b}\beta_j^2}{b-1} + n\sigma^2_{\alpha\beta}$
$\alpha\beta$	$(a-1)(b-1)$	$\sigma^2 + n\sigma^2_{\alpha\beta}$
ϵ	$ab(n-1)$	σ^2

Cuadro 3.1: ANOVA Mixto a dos Vías

En el caso del factor aleatorio la hipótesis nula es la siguiente:

$$H_0)\ \sigma^2_\alpha = 0 \tag{3.35}$$

$$H_0)\ \Sigma_\alpha = 0 \tag{3.36}$$

Donde la primera corresponde al caso univariado y la segunda al multivariado.

Mientas que el caso del factor fijo la hipótesis nula es la siguiente:

$$H_0)\ \beta_1 = \beta_2, \ldots, \beta_k = 0 \tag{3.37}$$

$$H_0)\ B_1 = B_2, \ldots, B_k = 0 \tag{3.38}$$

Donde la primera corresponde al caso univariado y la segunda al multivariado.

Al observar los cuadrados medios esperados del cuadro correspondiente al caso univariado se observa que;

- El estadístico F que permite testear la hipótesis nula del factor aleatorio es el cociente entre los cuadrados medios del propio factor y los cuadrados medios del error.
- El estadístico F que permite testear la hipótesis nula del factor fijo es el cociente entre los cuadrados medios de si mismo y los cuadrados medios de la interacción.

A continuación se detallan los estadísticos de Wilks y Pillai para cada una de las hipótesis anteriores:

- Hipótesis correspondiente al factor aleatorio.

$$\Lambda_\alpha = \frac{|H_{\alpha\beta}|}{|H_{\alpha\beta} + H_\beta|} \tag{3.39}$$

$$V_\alpha = tr\left[\left(H_{\alpha\beta} + H_\beta\right)^{-1} H_\alpha\right] \tag{3.40}$$

- Hipótesis correspondiente al factor fijo.

$$\Lambda_\beta = \frac{|E|}{|H + E_\beta|} \tag{3.41}$$

$$V_\beta = tr\left[\left(E + H_\beta\right)^{-1} H_\beta\right] \tag{3.42}$$

- Hipótesis correspondiente a la interacción.

$$\Lambda_{\alpha\beta} = \frac{|E|}{|H + E_{\alpha\beta}|} \tag{3.43}$$

$$V_{\alpha\beta} = tr\left[\left(E + H_{\alpha\beta}\right)^{-1} H_{\alpha\beta}\right] \tag{3.44}$$

Parte III

Resultados y conclusiones

Capítulo 4

Resultados

4.1. Datos utlizados

En una primera instancia se consideró pertinente relevar la totalidad de la población objetivo de la escuela 317 (Centro Comunal Zonal 6 de Montevideo), esto no fue posible debido a diversas causas. En un principio, la dificultad de contactar a cada uno de los individuos implicados en el estudio acarreó la pérdida de una pequeña proporción de la población (aproximadamente un 5 %). Luego, otra dificultad que se presentó fue el no cumplimiento en la entrega de la muestra de materia fecal a partir de la cuál se realizaría el correspondiente examen coproparasitario. Este problema significó la pérdida de un 50 % (aproximadamente) de los datos.

Parte del formulario relevaba información respecto de la ingesta de los niños y antecedentes de parasitosis. Esta parte de la encuesta debía ser respondida por los padres de los niños, adultos responsables o tutores para que los datos fueran válidos. La escasa concurrencia de adultos en los días de encuesta, motivó que los formularios fueran respondidos por los propios niños, lo que no es recomendable en menores de 9 años. Por esta razón estos datos no se consideraron.

De esta forma, los datos completos corresponden a un total de 104, entendiéndose por datos completos a aquellos casos en los que se cuenta con información antropométrica en ambas tomas y con información parasitaria pre intervención.

Para este estudio se tuvieron en cuenta las siguientes variables:

- Sexo: variable binaria.
- Variables antropométricas:
  - Talla: altura del individuo medida en centímetros.
  - Peso: medida en kilogramos.
  - IMC: índice de masa corporal, cociente entre el peso (kg) y la altura (mts) elevada al cuadrado.
- Variables ambientales:
  - Saneamiento.
  - Acceso al agua potable.
  - Variables derivadas del examen parasitario:
    - *Ascaris Lumbricoides*: variable binaria, indica la presencia/ausencia de este parásito.
    - *Trichuris Trichuria*: variable binaria, indica la presencia/ausencia de este parásito.
    - *Enterovius Vermicularis*: variable binaria, indica la presencia/ausencia de este parásito.
    - *Entamoeba Histolítica Dispar*: variable binaria, indica la presencia/ausencia de este parásito.
    - *Giardia Lamblia*: variable binaria, indica la presencia/ausencia de este parásito.
    - *Endamoeba Coli*: variable binaria, indica la presencia/ausencia de este parásito.
    - *Blastocystis Hominis*: variable binaria, indica la presencia/ausencia de este parásito.
    - *Iodamoeba Butschlii*: variable binaria, indica la presencia/ausencia de este parásito.
    - *Chilomastix Mesnili*: variable binaria, indica la presencia/ausencia de este parásito.
    - *Strongiloides Stercolaris*: variable binaria, indica la presencia/ausencia de este parásito.

- *Hymenolepis Nana*: variable binaria, indica la presencia/ausencia de este parásito.

A partir de las variables referidas al aspecto antropométrico de los individuos, se construyeron los llamados puntajes Z acorde a la metodología de la OMS [10]. Estas nuevas variables fueron, a su vez, discretizadas conformando así un nuevo juego de indicadores.

- Ind.Altura: variable ordinal que indica retraso severo en el crecimiento, retraso en el crecimiento y normal.
- Ind.Peso: variable ordinal que indica bajo peso severo, bajo peso y normal.
- Ind.IMC: variable ordinal que indica desnutrición grave, desnutrición moderada, normal, riesgo de sobrepeso, sobrepeso, obesidad.

Cabe aclarar que, salvo las variables ambientales, todas las demás fueron relevadas en ambas muestras. Así mismo, las variables derivadas del examen parasitario correspondiente a la segunda muestra, no fueron tenidas en cuenta debido a la extrema escasez de datos.

La estrategia global de análisis adoptada consistió en examinar los datos de cada una de las dos etapas del estudio mediante un análisis descriptivo y un análisis multivariado. El análisis multivariado corrsepondiente a la primera toma de datos comprendió los siguiente puntos:

- Estudiar la asociación entre los distintos tipos de parásitos mediante el gráfico de mosaico.
- Construir perfiles parasitarios utilizando la metodología de cluster probabilístico.
- Complementar los perfiles parasitarios con agrupaciones antropométricas. Validarlas utilizando MANOVA.
- Valerse del MANOVA para estudiar el impacto del saneamiento en el estado antropométrico de los individuos.

Por otro lado, luego del análisis exploratorio de la segunda toma de datos, el anlálisis multivariado siguió los siguientes puntos:

- Determinar si se produjo un cambio significativo en el estado parasitario de los niños valiéndose del test de homogeneidad marginal.
- Llevar a cabo la prueba T^2 de Hotelling para datos apareados sobre los Z scores para estudiar el cambio antropométrico entre las dos etapas

de la investigación.

- Investigar la influencia del saneamiento y los distintos parásitos en el cambio antropométrico de los individuos.

4.2. Primera toma de datos

4.2.1. Análisis descriptivo

Todos los cálculos llevados a cabo en este estudio fueron realizados en el software libre R [21]. A continuación se realiza una breve descripción de los datos utilizados en el presente trabajo.

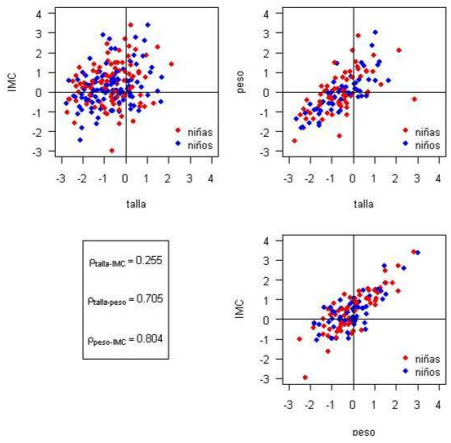

Figura 4.1: Z Scores Primera Toma de Datos Según Sexo

En la figura 4.1 se puede observar que existe una fuerte correlación entre los puntajes Z correspondientes a la talla y el peso, así como también, para el peso y el IMC.

		Niñas	Niños
Z talla	n	51	53
	media	-0,63	-0,634
	desvío	1,048	1,041
Z peso	n	37	33
	media	0,022	0,016
	desvío	1,396	0,873
Z IMC	n	51	53
	media	0,365	0,254
	desvío	1,299	1,046

Cuadro 4.1: Medidas de resumen por sexo para Z scores de peso, talla e IMC

Vale aclarar que la variable "Z peso" solo pudo calcularse para los individuos menores de diez años, debido a que la OMS considera que luego de esta edad la misma no es un buen indicador antropométrico. A partir del cuadro 4.1 y la figura 4.1 se puede pensar que no existe dimorfismo sexual en esta población, esto es, tanto los niños como las niñas presentan valores similares en los tres indicadores antropométricos. A través de cada uno de los puntajes Z, como se muestra en los cuadros 4.2 al 4.3, se calculó un indicador cualitativo, el cuál pretende resumir la información contenida en los Z scores. Los mismos suelen ser utilizados para realizar comparaciones con poblaciones de referencia.

Indicador	f. obs	p. obs	lim inf	lim sup	f. esp	p. esp
Z≤ -3	1	0,010	0,000	0,028	0	0,001
Z≤ -2	12	0,115	0,054	0,177	2	0,020
Z$>$ -2	91	0,875	0,811	0,939	102	0,979
Total	104	1,000	—	—		

Cuadro 4.2: Indicador de Altura

Indicador	f. obs	p. obs	lim inf	lim sup	f. esp	p. esp
Z ≤ -3	0	0,000	0,000	0,000	0	0,001
-3$\leq Z \leq -2$	1	0,015	0,000	0,045	1	0,020
2$\leq Z$	64	0,985	0,955	1,000	64	0,979
Total	65	1,000	—	—		

Cuadro 4.3: Indicador de Peso

Indicador	f. obs	p. obs	lim inf	lim sup	f. esp	p. esp
Z≤ -3	0	0,000	0,000	0,000	0	0,001
-3$\leq Z \leq -2$	1	0,010	0,000	0,028	2	0,020
-2$\leq Z \leq 1$	85	0,817	0,743	0,892	85	0,818
1$\leq Z \leq 2$	11	0,106	0,047	0,165	14	0,136
2$\leq Z \leq 3$	4	0,038	0,002	0,075	3	0,021
3$\leq Z$	3	0,029	-0,003	0,061	0,004	0
Total	104	1	—	—		

Cuadro 4.4: Indicador de IMC

En el cuadro 4.5 y la figura 4.2 se presentan los datos referentes a las variables ambientales. Puede verse como la gran mayoría de la población cuenta con agua potable pese a que solo la mitad de la población posee saneamiento.

		presenta	no presenta
Saneamiento	f.observada	45	49
	proporción	0,48	0,52
Agua Potable	f.observada	98	6
	proporción	0,94	0,06

Cuadro 4.5: Variables Ambientales

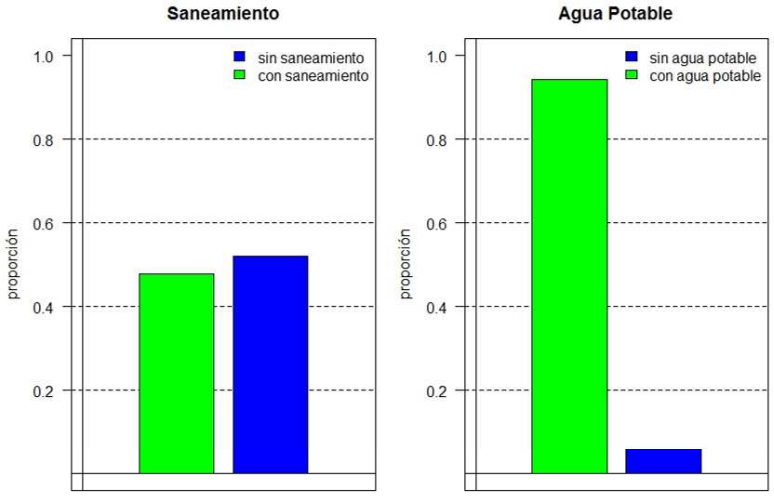

Figura 4.2: Variables Ambientales

Otro de los aspectos evaluados en esta muestra fue el referido a las variables coproparasitarias. En el cuadro 4.6 se muestran las frecuencias absolutas y relativas de los casos observados para cada una de las variables en consideración.

			presenta	no presenta
Geohelmintos	*Ascaris lumbricoides*	f.observada	38	41
		proporción	0,48	0,52
	Trichuris trichuria	f.observada	26	53
		proporción	0,33	0,67
	Hymenolepis nana	f.observada	2	77
		proporción	0,03	0,97
	Strongiloides stercolaris	f.observada	0	79
		proporción	0	1,00
Otros Patógenos	*Enterobius* vermicularis	f.observada	13	66
		proporción	0,16	0,84
	Endamoeba histolítica	f.observada	5	74
		proporción	0,06	0,94
	Giardia lamblia	f.observada	17	62
		proporción	0,22	0,78
Otros No Patógenos	*Entamoeba coli*	f.observada	18	61
		proporción	0,22	0,78
	Endolimax nana	f.observada	14	65
		proporción	0,18	0,82
	Blastocystis hominis	f.observada	6	73
		proporción	0,08	0,92
	Isospora belli	f.observada	0	79
		proporción	0	1,00
	Chilomastix mesnili	f.observada	0	79
		proporción	0	1,00

Cuadro 4.6: Variables Coproparasitarias - primera muestra

Se observa que los parásitos más prevalentes en esta población son Ascaris, Trichuris, Endamoeba Coli, Endolimax Nana, Giardia y Enterobius Vermicularis. La información del cuadro anterior, se presenta agregada en el siguiente cuadro. De esta manera, se observa la preponderancia de los Geohelmintos, siendo estos, los responsables de las infecciones más severas.

4.2.2. Análisis multivariante

Una primera aproximación al análisis de la asociación entre los tipos de parásitos se realiza gráficamente a través de un gráfico de mosaico. Según

[17] se trata básicamente es una visualización de áreas proporcionales a las frecuencias registradas en una tabla de contingencia. El mismo está compuesto de *azulejos* (correspondientes a cada celda de la tabla) creados recursivamente a través de particiones verticales y horizontales de un rectángulo. De esta forma, el área de cada rectángulo es proporcional a la celda correspondiente dadas las dimensiones de las particiones anteriores.

En los gráficos de mosaico, las dimensiones de las cajas representan la frecuencia relativa de individuos en cada combinación de categorías respecto al total, lo cual ofrece una aproximación gráfica de la distribución conjunta de las variables que componen el gráfico.

En la figura 4.3, se puede observar como el ancho y alto de las cajas no varían mucho en las diferentes combinaciones de tipos de parásitos. Por ejemplo (observando dos de las tres variables involucradas) se puede ver como la proporción de otros patógenos (división izquierda-derecha) no parece variar entre los individuos con o sin geohelmintos (primera división superior-inferior). Otro ejemplo sería, dentro de los niños que están infectados con geohelmintos (parte infereior del gráfico) la distribución de otros no patógenos no cambia mucho al considerar sujetos que tengan o no otros patógenos. El estadístico de Pearson asociado a la tabla (χ^2=1.81) que dio lugar a este gráfico confirma la sospecha de independencia entre estas variables.

4.2.2.1. Análisis de cluster - variables parasitarias

A partir de las variables descritas anteriormente se aplicó la metodología de cluster, con el fin de definir perfiles parasitarios de la población en cuestión. Para ellos se utilizó la metodología de análisis de cluster basado en un modelo probabilístico, expuesta por Moustaki Papageorgiou [18]. El código utilizado para llevar a cabo esta metodología se encuentra en el anexo A.1.1. El criterio utilizado para seleccionar el número de grupos fue el BIC.

	τ	ϕ_{geo}	ϕ_{op}	ϕ_{nop}
Grupo 1	0,36	0,55	0,49	0,81
Grupo 2	0,26	0,96	0,34	0,07
Grupo 3	0,38	0,25	0,33	0,12

Cuadro 4.7: Parámetros Estimados para los tres Grupos

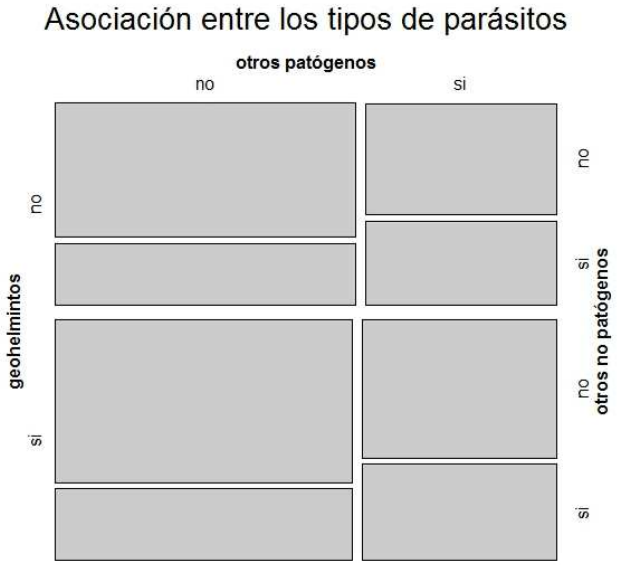

Figura 4.3: Gráfico de mosaico para distintos tipos de parásitos

En el cuadro 4.7 se presentan tanto las probabilidades a priori (τ_i) de cada uno de los tres grupos, así como la probabilidad de estar infectado con cada uno de los distintos parásitos (ϕ_{jk}). Estas probabilidades a priori describen la proporción de cada una de las tres distribuciones conjuntas en la mezcla global. Una vez caracterizados los grupos, las probabilidades a priori también pueden ser interpretados como las probabilidades de que un cierto individuo padezca las infecciones antes de observar su perfil parasitario. Para tener una mejor visualización se presenta el gráfico 4.4.

El tamaño relativo de cada circunferencia responde a la probabilidad a priori asignada a cada grupo. Dentro de cada una hay tres sectores, uno para cada una de las tres variables. Por último, cada sector es coloreado de acuerdo a la proporción estimada de individuos que poseen cada infección dentro de ese grupo. A modo de ejemplo, véase el sector "nop" (no patógenos) del primer grupo (arriba a la izquierda). Allí se puede ver que los niños del primer grupo poseen una alta probabilidad de padecer infecciones provocadas por otros no patógenos. Las líneas punteadas representan los porcentajes 25, 50 y 75 por ciento contraer las infecciones. De aquí se desprende que en el grupo dos la probabilidad de contrar infecciones causadas por geohelmintos es muy elevada. Los individuos del grupo tres poseen una reducida probabilidad de

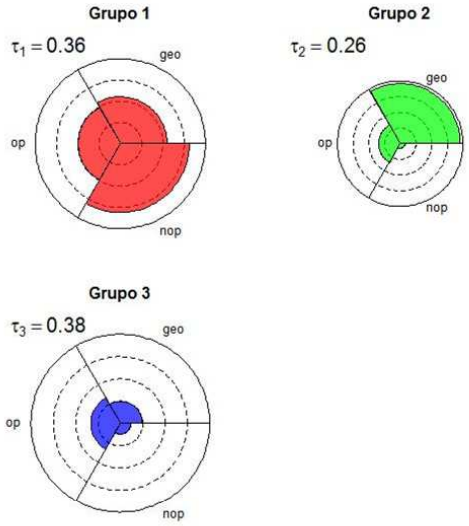

Figura 4.4: Grupos parasitarios

contraer infecciones parasitarias. Por otro lado, los individuos pertenecientes al grupo uno son los más propensos a contraer cualquiera de las infecciones, principalmente las provocadas por el grupo de los no patógenos.

El criterio utilizado para seleccionar el número de grupos fue el BIC. Como se observa en el gráfico 4.5, el número de grupos con BIC mínimo fue tres, alcanzando un valor de 1294,24.

En el cuadro 4.8 se procedió a validar los resultados comparando las frecuencias observadas con las respectivas frecuencias esperadas asumiendo que el número de grupos es el correcto.

		No patógenos			
Geohelmintos	**Otros patógenos**	no		si	
no	no	15	(14,99)	7	(7.03)
	si	8	(8,00)	6	(5,96)
si	no	18	(18,00)	8	(7,97)
	si	10	(9,99)	7	(7.03)

Cuadro 4.8: Frecuencias Observadas y Esperadas en los Grupos Parasitarios

La información presentada en paréntesis corresponde a las frecuencias esperadas, mientras que la dispuesta sin paréntesis representa las frecuencias efectivamente observadas. El estadístico de Pearson asociado a esta tabla es

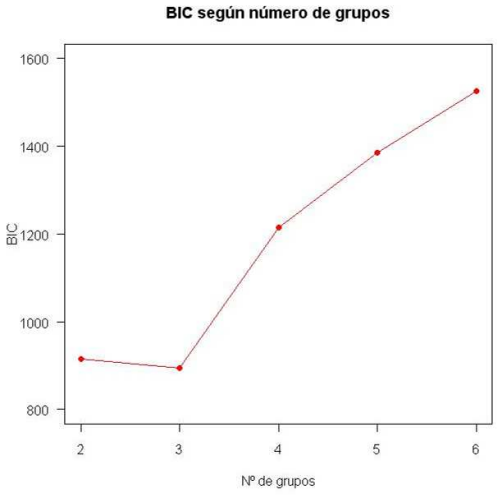

Figura 4.5: Criterio de selección del número de grupos

de 6,2x10^{-4} con un p-valor asociado de 0,98, lo que indica que este modelo describe adecuadamente los datos.

4.2.2.2. Análisis de cluster - variables antropométricas

Así como se generaron tres grupos con las variables coproparasitarias, también se generaron los correspondientes grupos para las variables Z scores. El método utilizado para lograr dicho propósito fue el de Ward. Luego, al analizar los indicadores R^2 y pseudo-F, se tomó la decisión de trabajar con tres grupos, como indica el cuadro: 4.9

Grupos	R^2	psF
5	0,745	54,054
4	0,696	57,179
3	0,632	65,258
2	0,391	49,357
1	0	–

Cuadro 4.9: Cantidad de Grupos

Para caracterizar los grupos se hizo uso de análisis de los gráficos de caja que se expone en la figura 4.6

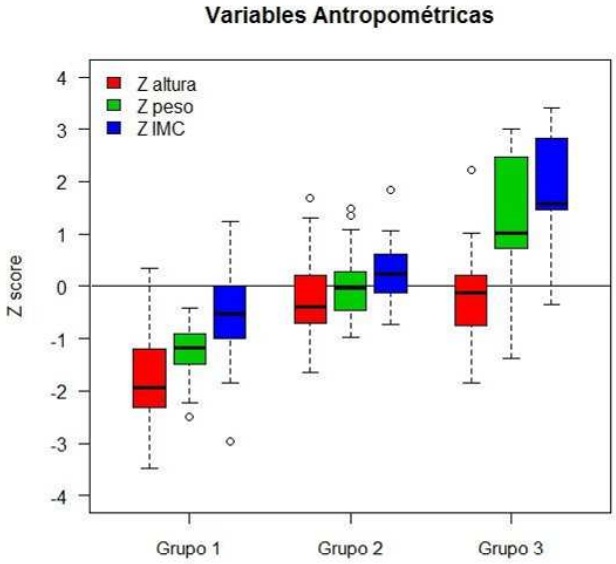

Figura 4.6: Variables Antropométricas

Los grupos conformados de esta manera pueden caracterizarse de la siguiente forma. Los niños incluidos en el grupo uno son aquellos cuyos indicadores antropométricos se encuentran más comprometidos. Por otro lado, los que integran el grupo tres se encuentran en la situación opuesta, mientras tanto los individuos del segundo grupo están en un punto intermedio. Para respaldar esta caracterización se puso en práctica un análisis multivariado de la varianza (MANOVA).

4.2.2.3. Inferencias - primera muestra

		Z altura	Z IMC	Z peso
Grupos antropométricos	grupo 1	-1,648	-0,526	-1,175
	grupo 2	-0,234	0,256	0,013
	grupo 3	-0,166	2,185	1,566

Cuadro 4.10: Vectores de Medias de los Grupos Antropométricos

	Est.Pillai	Aprox.F	glH	glE	p-valor
Grupos antropométricos	0,803	14,778	6	132	$<1{,}0\text{x}10^{-7}$

Cuadro 4.11: MANOVA para IMC, Peso y Altutra Iniciales

A través del análisis que se presenta en el cuadro 4.11 se pudo concluir que este agrupamiento produjo diferencias significativas en los vectores de medias de los tres Z-scores.

La siguiente etapa en el análisis de estos datos consistió en determinar si existía algún tipo de asociación entre los dos métodos de agrupación descriptos hasta aquí.

		Grupos Parasitarios		
		grupo 1	grupo 2	grupo 3
Grupos Antropométricos	grupo 1	9	6	7
	grupo 2	17	16	12
	grupo 3	2	6	4

Cuadro 4.12: Grupos Parasitarios

		Grupos Parasitarios		
		grupo 1	grupo 2	grupo 3
Grupos Antropométricos	grupo 1	0,321	0,214	0,304
	grupo 2	0,607	0,571	0,522
	grupo 3	0.071	0,214	0,174

Cuadro 4.13: Perfiles Columna- Grupos Parasitarios

El estadístico χ^2 correspondiente al cuadro 4.12 es de 2.81, por lo tanto parece razonable pensar que las variables estudiadas son independientes. Pese a que en primera instancia podría haberse creído que cierto tipo de infecciones desfavorecen el crecimiento más que otras (lo cual se puede apreciar en los perfiles columna), la evidencia brindada por este resultado no apoya dicha teoría.

Un resultado interesante en el análisis de esta primera toma de datos es el que se muestra en las tablas 4.14 y 4.15 respectivamente:

		grupo 1	grupo 2	grupo 3
Saneamiento	No	10	16	7
	Si	16	7	14

Cuadro 4.14: Grupos Parasitarios según tenencia de Saneamiento

		grupo 1	grupo 2	grupo 3
Saneamiento	No	0,303	0,485	0,212
	Si	0,432	0,189	0,378

Cuadro 4.15: Perfiles fila-Grupos Parasitarios según tenencia de Saneamiento

Se puede ver como los niños del grupo tres, que son los que tienen menos probabilidad de estar infectados, poseen en su mayoría saneamiento. El p-valor del estadístico de Pearson asociado a esta tabla es de 0.03, lo cual respalda la suposición de asociación entre el estado sanitario de la vivienda de los niños y la presencia de parásitos.

Al realizar el análisis de la separación de los vectores de medias de los Z-scores en la presencia o ausencia de saneamiento, cuadro 4.16, el resultado obtenido mediante el estadístico T^2 de Hotelling indica que los niños sin saneamiento están en condiciones significativamente inferiores a los niños que si cuentan con él.

		Z altura	Z IMC
Saneamiento	No	-0,73	0,39
	Si	-0,49	0,27

Cuadro 4.16: Vectores de Medias de Altura e IMC según tenencia de Saneamiento

La aproximación F del estadístico T^2 de Hotelling y su p-valor fueron de 8.475 y 5.0x10^{-4} respectivamente. Una vez rechazada la hipótesis de igualdad de los vectores de medias, se procedió a analizar los coeficientes de la función discriminante como se muestra en el cuadro 4.17.

Z altura	Z IMC
-1,689	0,067

Cuadro 4.17: Función Discriminante

Se puede ver como la variable que contribuye de mayor forma a la separación de los vectores de medias es la altura.

4.2.3. Síntesis de la primer toma de datos:

En cuanto al apartado parasitario, se logró caracterizar la población en los siguientes perfiles:

- Grupo 1: infecciones causadas por geohelmintos, otros patógenos y otros no patógenos.
- Grupo 2: infecciones causadas mayormente por geohelmintos.
- Grupo 3: individuos con una menor probabilidad de contraer cualquiera de las infecciones.

Asimismo, luego de inspeccionar los datos antropométricos surgieron tres grupos, los cuales se definieron así:

1. Grupo 1: individuos con valores bajos en los puntajes Z.
2. Grupo 2: individuos con valores intermedios en los puntajes Z.
3. Grupo 3: individuos con valores altos en los puntajes Z.

Al analizar la tabla de contingencia resultante de cruzar las dos agrupaciones anteriores, se llegó a la conclusión de que ambas no evidencian asociación alguna. Lo cual resulta llamativo debido a que podría esperarse que niños con un perfil parasitario más comprometido debieran presentar Z scores más bajos y viceversa.
Por último, se observó que la presencia de saneamiento jugó un papel importante tanto en el estado antropométrico como parasitario de los niños.

4.3. Segunda toma de datos

4.3.1. Análisis descriptivo

A continuación se realiza una breve descripción de los datos correspondientes a esta etapa del estudio:

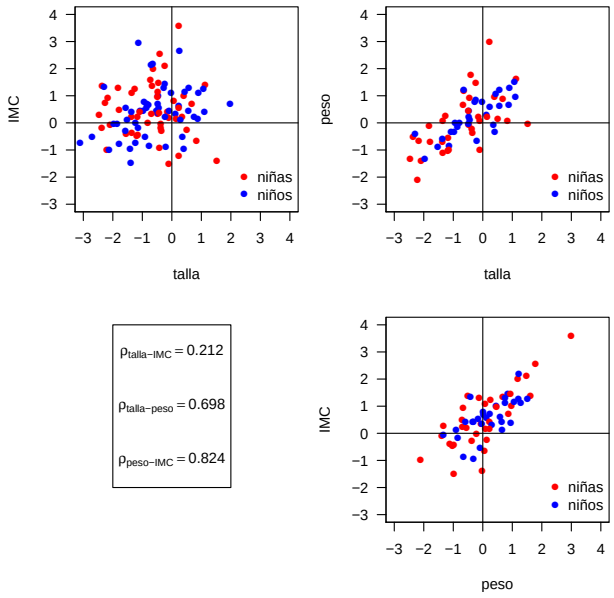

Figura 4.7: Z Scores Segunda Toma de Datos Según Sexo

De esta forma se puede observar que existe una fuerte correlación entre los puntajes Z correspondientes a la talla y el peso, así como también, para el peso y el IMC como ya se había observado en la primera toma de datos.
Para evaluar la evolución antropométrica de los individuos se presenta en el gráfico 4.8

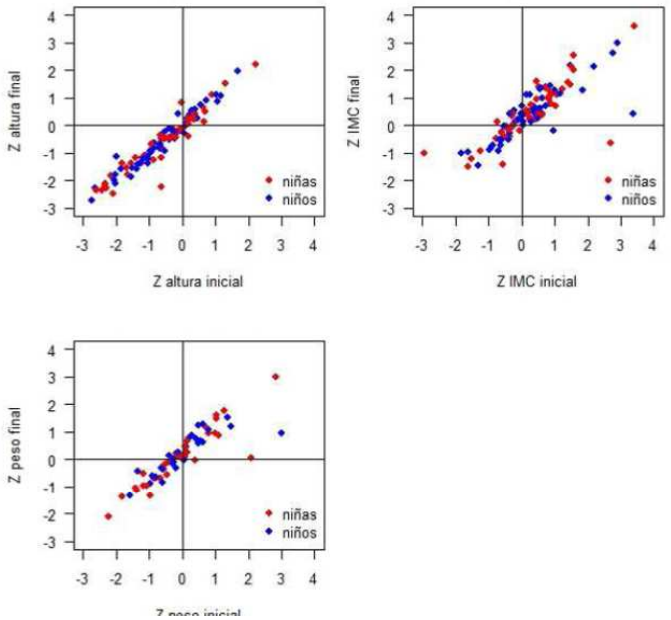

Figura 4.8: Z Scores

Al igual que en la primer toma de datos, se calcularon los siguientes indicadores antropométricos de carácter cualitativo, cuadro 4.18, cuadro 4.19 y cuadro 4.20. Los mismos serán utilizados junto a sus similares de la primera fase del estudio para evaluar el impacto de la intervención en el estado de los niños.

Indicador	Cantidad	Proporción	lim inf	lim sup	p. esperado	f. esperada
$\leq -3Z$	2	0,019	0,000	0,046	0,001	0,000
$\leq -2Z$	10	0,096	0,039	0,153	0,020	0,000
$>$ -2 Z	92	0,885	0,823	0,946	0,979	0,000
Total	104	1,000	—	—		

Cuadro 4.18: Altura para la Edad Final

Indicador	Cantidad	Proporción	lim inf	lim sup	p. esperado	f. esperada
$\leq -3Z$	0	0,000	0,000	0,000	0,001	0,061
$\leq -2Z$	1	0,015	0,000	0,045	0,020	1,220
$>$ 2 Z	64	0,985	0,955	1,000	0,979	59,719
Total	65	1,000	—	—		

Cuadro 4.19: Peso para la Edad Final

Indicador	Cantidad	Proporción	lim inf	lim sup	p. esperado	f. esperada
$\leq -3Z$	0	0,000	0,000	0,000	0,001	0,079
$>$ -3 Z y $\leq -2Z$	0	0,000	0,000	0,000	0,020	1,580
$>$ -2 Z y $\leq 1Z$	76	0,731	0,646	0,816	0,818	64,622
$>$ 1 Z y $\leq 2Z$	20	0,192	0,117	0,268	0,136	10,744
$>$ 2 Z y $\leq 3Z$	6	0,058	0,013	0,103	0,021	1,659
$>$ 3 Z	2	0,019	0,000	0,046	0,004	0,316
Total	104	1	—	—		

Cuadro 4.20: IMC para la Edad Final

Al procesar la información de las variables parasitarias correspondientes a esta instancia, se encontraron los resultados que muestra el cuadro 4.21

		presenta	no presenta
Geohelmintos	f.observada	7	23
	proporción	0,23	0,77
Otros	f.observada	7	23
Patógenos	proporción	0,23	0,77
Otros	f.observada	0	30
No Patógenos	proporción	0	1,00

Cuadro 4.21: Variables Coproparasitarias - segunda Muestra

El inconveniente principal de esta etapa fue la dificultad para recolectar las muestras necesarias para el análisis coproparasitario. Solo fue posible cubrir el 28,8 % de los 104 individuos estudiados.

4.3.2. Análisis multivariante

Pese a que en esta segunda toma de datos (30 exámenes parasitarios) la proporción de niños infectados se redujo sustancialmente con respecto a la primera (104 exámenes parasitarios), al analizar los exámenes pre y post intervención de los 30 datos, se observa que esta reducción no es significativa. Esto se debe a que en principio los 30 individuos se encontraban en condiciones ligeramente mejores en relación al resto de la muestra. Al comparar el estado inicial y final de los treinta niños que completaron esta etapa mediante el test de McNemar (test de homogenidad marginal), se observan los siguientes resultados (cuadro 4.22).

			Después	
			no presenta	presenta
Antes	Geohelmintos	no presenta	16	2
		presenta	7	5
	Otros Patógenos	no presenta	14	2
		presenta	9	5
	Otros no Patógenos	no presenta	18	0
		presenta	12	0

Cuadro 4.22: Variación en los Indicadores Parasitarios

Dado que no es correcto considerar como independientes el antes y el después del mismo niño, estos datos fueron tratados como apareados. Por ende el contraste indicado para analizar estas tablas de 2×2 es la prueba de McNemar. El resultado de este estadístico (cuadro 4.23) para las dos primeras tablas indica que no se produjo un cambio significativo entre el antes y el después de estos individuos.

	Estadístico	p-valor
Geohelmintos	1,78	0,18
Otros Patógenos	3,27	0,07
Otros No Patógenos	-	-

Cuadro 4.23: Contraste de McNemar

El siguiente paso de la investigación consiste en analizar si existe o no un cambio en las variables antropométricas de los individuos. Dado que la altura, el peso y el IMC describen de manera conjunta el problema a estudiar,

se optó por contrastar de forma multivariada a las variables mencionadas anteriormente.

Aquí se presenta el cuadro 4.24 que resume la variación de los puntajes Z entre las dos fases del estudio:

		Antes	Después	Diferencia
Z talla	n	104	104	104
	media	-0,632	-0,631	0,001
	desvío	1,039	1,066	0,321
Z peso	n	70	65	65
	media	0,019	0,201	0,182
	desvío	1,171	1,106	0,474
Z IMC	n	104	104	104
	media	0,309	0,481	0,172
	desvío	1,172	1,073	0,627

Cuadro 4.24: Diferencias Z scores

Se aprecia una leve mejoría en los tres aspectos relevados. A continuación se presenta el resultado del contraste de Hotelling para datos apareados (cuadro 4.25). Dada la escasez de datos en el peso de los niños, se optó por llevar a cabo el contraste incluyendo solo la altura y el IMC por un lado e incluyendo las tres variables por otro.

Estadístico T^2	N^o de variables	p-valor
6,748	2	0,006
7,607	3	0,071

Cuadro 4.25: Contraste de Hotelling

Pese a que en el caso en el que se consideran las tres variables, el p-valor es levemente superior al 5 %, se optó por rechazar la hipótesis nula en ambos casos, esta postulaba que el vector de medias de las diferencias era el nulo. Al rechazar dicha hipótesis, se llevaron a cabo las pruebas univariadas "protegidas" (cuadro 4.26). Los resultados indican que pese a que la altura no sufre cambios significativos entre muestras, el peso y el IMC aumentan ligeramente debido a la intervención.

	Estadístico	p-valor
Z altura	-0,031	0,976
Z IMC	-2,800	0,006
Z peso	-2,570	0,012

Cuadro 4.26: Pruebas t Apareadas Protegidas

Como alternativa a este enfoque, se planteó analizar el cambio antropométrico medido a través de los indicadores cualitativos construidos a partir de los puntajes Z. En primer lugar se presentan las tablas contingencia que comparan dichas variables (cuadro 4.27, cuadro 4.28 y cuadro 4.29).

		después			
		normal	retraso	retraso severo	total
antes	normal	89	1	1	91
	retraso	3	9	0	12
	retraso severo	0	0	1	1
	total	92	10	2	104

Cuadro 4.27: Cambios en la Altura

		después			
		normal	bajo peso	bajo peso severo	total
antes	normal	64	0	0	64
	bajo peso	0	1	0	1
	bajo peso severo	0	0	0	0
	total	64	1	0	65

Cuadro 4.28: Cambios en el Peso

		después			
		bajo IMC	normal	alto IMC	total
antes	bajo IMC	0	1	0	1
	normal	0	72	13	85
	alto IMC	0	3	15	18
	total	0	76	28	104

Cuadro 4.29: Cambios en el IMC

El estadístico de homogeneidad marginal (calculado mediante la librería coin[13] del R) para este caso proporciona un valor de 0,37 por lo cual no se rechaza la hipótesis nula, o lo que equivale a decir que no se produjeron cambios entre las muestras. En el caso del peso (cuadro 4.28) no fue necesario llevar a cabo el test, ya que al observar la tabla correspondiente no se observa ningún cambio entre muestras. Distinto es el caso del indicador correspondiente el IMC. En este caso, el estadístico de homogeneidad marginal favorece a la hipótesis alternativa (p-valor= 0,02). Por lo cual se puede concluir que se produjeron cambios en el IMC de los niños respecto de ambas muestras. Los resultados correspondientes a la altura e IMC confirman aquellos obtenidos mediante la prueba de Hotelling. Sin embargo, debe tenerse en cuenta que este tipo de indicadores no retienen toda la información de los Z scores, ya que los mismos son una discretización de estos. No obstante, los valores obtenidos tras los contrastes de homogeneidad marginal confirman que el cambio en el IMC es suficiente como para ser captado tanto a través de las variables cuantitativas así como de las cualitativas.

En último lugar se analizó si el cambio antropométrico dependió de la ausencia de saneamiento, si fue diferente en los distintos grupos parasitarios o si obedeció al sexo de los individuos. Los cálculos fueron realizados utilizando la función *lmer* de la librería lme4[4] del R. Para ello se planteó el siguiente modelo multivariante:

$$Y_{ijklm} = \mu + Ind_i + San_j + Sexo_k + Grupo_l + \epsilon_{ijklm} \tag{4.1}$$

$$i = 1, 2 \ldots, n \quad j = 1, 2 \quad k = 1, 2 \quad l = 1, 2, 3 \quad m = 1, 2 \ldots, n_{ijkl}$$

$$\epsilon_{ijklm} \sim N_6\left(0, \Sigma\right)$$

$$Ind_i \sim N_6\left(0, \Sigma_{Ind}\right)$$

Mediante este modelo se buscó investigar si existió o no un cambio en los puntajes Z individuales en los distintos niveles de saneamiento, sexo y grupos parasitarios. En el cuadro 4.30 se detallan los resultados de ajustar este modelo.

	Est.Pillai	Aprox.F	glH	glE	p-valor
Individuo	0,064	0,973	3	43	0,414
Saneamiento (individuo)	0,163	2,784	3	43	0,052
Sexo (individuo)	0,025	0,363	3	43	0,779
Grupo (individuo)	0,074	1,153	3	43	0,338

Cuadro 4.30: Manova para Mediciones Repetidas- Altura IMC Peso

De esta forma se puede apreciar como el hecho de tener o no saneamiento repercutió en el cambio de los Z scores. Para ilustrar esto se presenta el siguiente gráfico.

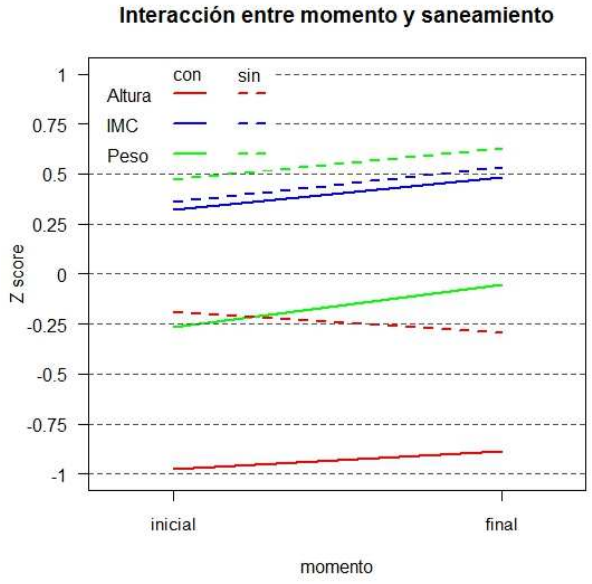

Figura 4.9: Manova de los Z scores respecto a los momentos y al saneamiento

Queda claro como la manera en la que la altura de los individuos cambió entre muestras es diferente según la tenencia o no de saneamiento, esto se ve dado que las lineas rojas (completa y punteada) no son paralelas.

En el caso de que no se tome en cuenta el peso, los resultados obtenidos son los presentados en el cuadro 4.31.

	Est.Pillai	Aprox.F	glH	glE	p-valor
Individuo	0,049	1,661	2	65	0,198
Saneamiento (individuo)	0,120	4,437	2	65	0,016
Sexo (individuo)	0,037	1,239	2	65	0,296
Grupo (individuo)	0,027	0,905	2	65	0,410

Cuadro 4.31: Manova para Mediciones Repetidas- Altura, IMC

Al igual que en el caso anterior, se aprecia como el saneamiento es un factor determinante. Otro de los productos que se pudieron obtener de estos análisis fueron las medidas de asociación entre las variables. Los mismos se detallan en el cuadro 4.32.

	Modelo con Z score Peso		Modelo sin Z score Peso	
	η^2 Wilks	η^2 Roy	η^2 Wilks	η^2 Roy
Saneamiento	0,163	0,163	0,121	0,120
Grupo parasitario	0,075	0,074	0,037	0,036
Sexo	0,025	0,025	0,027	0,027

Cuadro 4.32: Asociación entre las Variables

Puede observarse como, en ambos modelos, la única variable que presenta una cierta asociación con los puntajes Z es el saneamiento.

Vale aclarar, que para ambos modelos (con y sin Z score correspondiente al peso) se llevó a cabo una etapa de diagnóstico que validara los supuestos requeridos. Para chequear el supuesto de homogeneidad de matrices de covarianzas se utilizó el estadístico Box-M. Los resultados se detallan en las tablas 4.33 y 4.34.

	Estadístico M	aprox.F	p-valor
Saneamiento	93,804	0,899	0,593
Sexo	83,226	0,798	0,702
Grupo parasitario	90,970	0,872	0,622

Cuadro 4.33: Estadístico Box-M (Z score con peso)

Se ve que, en ambos modelos, este supuesto no es violado para ninguna de las variables. En el cuadro 4.35 se observa que no se cumple el supuesto de multinormalidad en ambos modelos. La prueba fué realizada con el código

	Estadístico M	aprox.F	p-valor
Saneamiento	47,702	0,999	0,488
Sexo	36,996	0,698	0,711
Grupo parasitario	27,507	0,576	0,808

Cuadro 4.34: Estadístico Box-M (Z score sin peso)

que se explicita en el anexo A.2. Sin embargo gracias a la robustez del estadístico de Pillai frente al no cumplimiento de este supuesto, los resultados siguen teniendo validez.

	Estadístico	gl	p-valor
Modelo con Z score peso	64,299	8	$<1{,}0\times10^{-7}$
Modelo sin Z score peso	136,772	12	$<1{,}0\times10^{-7}$

Cuadro 4.35: Estadístico Doornik-Hansen

4.3.3. Síntesis de la segunda toma de datos:

De la misma forma que en la primera toma de datos, esta fase del estudio estuvo marcada por la falta de muestras coproparasitarias. De esta manera solo se pudieron complementar los análisis de laboratorio iniciales de treinta niños. Recabados estos datos, se procedió a comparar la presencia/ausencia de los tres principales grupos de parásitos en los dos momentos del tiempo. La conclusión a la que se llegó fue que, pese a que se llevó a cabo una intervención donde los niños fueron adecuadamente medicados, promedialmente no se registraron grandes cambios. Visto esto, el equipo del Instituto de Higiene y Escuela de Nutrición propuso la explicación de que esta situación pudo generarse debido a que estos niños viven en un ambiente muy carenciado y están en contacto constante con los agentes transmisores de los patógenos, por lo cual, luego de estar libres de parásitos durante un tiempo, volvieron a recaer. Sin embargo, el tiempo en el que estuvieron libres de parásitos fue suficiente para que, en promedio, se diera un leve cambio significativo en las variables que resumen el estado antropométrico de esta población. Al mismo tiempo, mediante el estadístico de Hotelling se concluyó que los Z scores variaron significativamente entre muestras, siendo el IMC la variable de mayor cambio. En última instancia se trató de investigar cuales fueron los factores que favorecieron o no el cambio antes mencionado. Es así que, utilizando modelos multivariantes mixtos se pudo observar que el hecho de tener o no

saneamiento fue un componente determinante para el cambio antropométrico.

Capítulo 5

Juicios finales

5.1. Conclusiones

Todas las conjeturas presentadas en este apartado están influenciadas por la falta de información, la cual fué una de las principales limitantes del estudio realizado, las mismas se debieron a varias causas. Por un lado se dió el abandono temprano de algunos de los estudiantes que concurrían a la escuela en cuestión y por otro, el seguimiento de los demás resultó sumamente complejo dada la delicada y precaria situación en la que los niños estaban inmersos. Por si fuera poco, una complicación extra fue que durante las entrevistas en las que se planeaba relevar el apartado nutricional, los antecedentes familiares de parasitosis y algunos puntos más, los padres/abuelos/tutores de los niños no se encontraban presentes, lo cual llevó a que estos ítems del cuestionario no tuvieran la seriedad deseada y por ende, su descripción y análisis no se incluyera en este estudio. A su vez, dentro de los datos que sí se pudieron recolectar, la mayor limitante estuvo presente en la parte referente a las variables parasitarias, ya que solo se logró obtener una tasa de respuesta muy baja de la población (alrededor de 50 % en la primera instancia y 15 % en la segunda). Esto llevó a tomar la decisión de examinar 104 niños, en tanto esa fue la cantidad para la cual se dispuso de datos antropométricos completos en ambas instancias y datos parasitarios al menos en la primera.

5.1.1. Primer toma de datos

A modo de partida, en primer lugar se investigó si la población presentaba algún tipo de dimorfismo sexual o diferencias significativas en cuanto a la edad a tener en cuenta para las tres variables antropométricas. La hipótesis formulada en torno al dimorfismo sexual fué descartada utilizando el estadístico t de Student, de esta manera se concluyó que al principio del estudio, niños y niñas presentaban similares características en cuanto a talla, IMC y peso se refiere. La edad no presentó problemas ya que las variables antropométricas fueron estandarizadas (convertidas en *Z scores*) mediante el procedimiento de la OMS descrito en la sección 3.1.

La manera de abordar el objetivo referente a la creación de tipologías de individuos fue la siguiente: durante la primera fase de la evaluación se crearon (a partir de los datos mismos) diferentes tipos de agrupamientos. Los dos enfoques utilizados tuvieron en cuenta, por un lado, la información antropométrica y por otro la correspondiente a la parte parasitaria. Utilizando los indicadores descritos en la sección 3.7.1, en ambos casos se determinó que el número óptimo de grupos era tres. Fue así que, los grupos antropométricos resultantes marcaron tres categorías, la existencia de un conjunto de individuos en condiciones "normales" y dos grupos en situaciones opuestas, uno de ellos con niveles altos en los puntajes Z y el otro en la situación contraria, siendo este último el más comprometido de los tres. Por otro lado, los grupos construidos a partir de la información parasitaria revelaron la existencia de una cohorte conformada por niños propensos a cualquiera de las infecciones, otra compuesta por niños principalmente afectados por geohelmintos y un último grupo compuesto por individuos con bajas probabilidades de infección.

Una vez finalizada la primera instancia de evaluación se planteó la hipótesis natural de investigar si existía algún tipo de asociación entre las dos tipologías construídas. Mediante el uso del estadístico χ^2 de Pearson se llegó a la conclusión de que las tipologías antropométricas y parasitarias no presentan asociación alguna, siendo esto curioso en tanto a que podría esperarse que niños más afectados por los agentes parasitarios deberían tener cierto retraso en la talla y/o IMC. Adicionalmente se trabajó sobre el impacto del saneamiento en la situación de los sujetos. Fue así que, mediante el cálculo del estadístico T^2 de Hotelling, se vió que los puntajes Z de la primera toma de datos presentaban diferencias significativas acorde a la presencia o ausencia

de saneamiento. Por otro lado, se usó el estadístico χ^2 de Pearson para comprobar si el saneamiento era un factor condicionante a la hora de evaluar el estado parasitario de los individuos. De esta manera se determinó que aquellos niños cuyos hogares disponen de saneamiento se encuentran en mejores condiciones tanto en el estado antropométrico como parasitario.

5.1.2. Intervención

Luego de aproximadamente un mes de iniciado el estudio, se recolectaron los resultados de las espátulas adhesivas y las muestras fecales. Los resultados brindados por estos exámenes permitieron determinar que medicamento debía ser suministrado a cada niño. Luego de haber pasado cinco meses de esta fase de intervención, en la que los niños infectados recibieron medicamentos acorde a sus necesidades, se volvierion a relevar los componentes claves del estudio (puntajes Z y presencia de parásitos).

5.1.3. Segunda toma de datos

En esta segunda toma de datos se observó como las altas correlaciones observadas *entre* los indicadores antropométricos se mantenían. Luego, aprovechando que se disponía de información para ambas instancias, se consideraron los momentos "anterior" y "posterior" a la intervención para observar correlaciones *dentro* de las variables. Fue así que también se vieron altas correlaciones entre el antes y después de cada variable. Indicando que, en principio, niños en situaciones comprometidas de talla, peso e IMC tendían a permanecer en esa situación, lo mismo que los niños en situación de sobrepeso.

Por otro lado se trató de cuantificar la relación temporal entre las variables parasitarias (presencia o ausencia de cada patógeno), pero la escasez de muestras disponibles (sólo 30 casos de 104 iniciales, lo cual comprende tan solo el 28,8 % de la cantidad inicial) no permitió obtener descripciones tan precisas como se hubiese querido. Debido a esta escasez de datos, se utilizaron procedimientos de índole no paramétrica. De esta manera, a través del contraste de McNemar se identificó un cambio significativo *entre* muestras en la proporción de individuos que padecían de parásitos no patógenos, cambio que no se registró en los geohelmintos ni en los otros patógenos. A la vista de

estos resultados, se concluyó que la intervención no fue 100 % efectiva debido a que dadas las precarias condiciones del lugar donde viven los niños, estos están en permanente contacto con los agentes transmisores de los parásitos.

Disponiendo de las variables antropométricas de las dos fases de la investigación, se seleccionaron métodos multivariados de carácter longitudinal para su análisis. Dado que solo se disponía de dos instancias temporales, se optó por representar el cambio en estas variables a través de los incrementos de las mismas, esto es, la diferencia entre el valor post-intervención y el pre-intervención. Considerando los resultados proporcionados por el estadístico T^2 de Hotelling para datos apareados, se llegó a la conclusión de que si bien se produjo un cambio en el vector de medias en dicho conjunto de variables, este se manifestó de una forma muy leve. Habiendo rechazado la hipótesis nula (multivariada) de esta prueba se procedió a realizar las pruebas t "protegidas" con el fin de investigar cual o cuales de las tres variables antropométricas presentaron el mayor cambio entre muestras. De esta manera se vió que los puntajes Z que presentaron mayor influencia sobre el cambio antropométrico fueron el peso y el IMC. Esto muestra que, pese a que los niños solo estuvieron libres de parásitos por un breve período de tiempo, este fue suficiente para que presentaran una mejoría leve. Cabe señalar que dicha mejora no se debe al paso del tiempo ya que el mismo es anulado al re-estandarizar los datos correspondientes a la talla, peso e IMC.

Este enfoque fue complementado a través del uso de los indicadores cualitativos construídos a partir la discretización de los Z scores. Mediante el uso del contraste de homogeneidad marginal (con la variante de Bhapkar) se replicaron las conclusiones ya obtenidas a través la prueba de Hotelling, es decir tanto el peso como el IMC variaron entre las instancias.

Por último se dispuso de un análisis multivariado de la varianza (MANOVA) para identificar las fuentes de variación del cambio antropométrico. Dicho análisis se valió tanto de factores fijos como aleatorios. El factor aleatorio al que se hace referencia son los propios niños ya que, como no se logró contar con todos los individuos de la escuela, se los consideró como una muestra aleatoria de la misma (aunque se desconce el mecanismo mediante el cal fue generada). Los factores considerados como fijos fueron el saneamiento, el sexo y los grupos parasitarios y antropométricos construidos en la primera toma de datos. Gracias a este modelo se apreció como el saneamiento no solo afectó la condición antropométrica inicial de los niños sino que también repercutió sobre la forma en que se produjeron los cambios entre los momentos.

Las variables que no resultaron significativas en este modelo fueron el sexo, los grupos parasitarios y antropométricos construídos en la primera toma de datos. Sobre la no significación de esta última (grupos antropométricos) puede concluirse que la variación antropométrica de los niños en el período no fue afectada por las condiciones iniciales de los mismos.

5.2. Consideraciones a futuro

Como consideraciones para futuras investigaciones se sugiere realizar la investigación controlando la evolución de las infecciones con mayor frecuencia y en períodos más reducidos de tiempo. Con esta metodología se podría lograr un mejor seguimiento del efecto de la ventana de tiempo en que la medicación hace efecto y del impacto de la misma. Por otra parte el equipo del Instituto de Higiene sugirió llevar a cabo una próxima investigación en niños cursando preescolar ya que en ellos se pueden apreciar mayores cambios antropométricos en un período de tiempo menos prolongado.

Bibliografía

[1] Acuña, M., Calegari, L., Dinder, C., Rosa, R., Salvatella, R., Savio, M., Castell, R., and Zanetta, E. Helmintiasis intestianal, manejo de geohelmintiasis., Enero 2003.

[2] Acuña, M., Da Rosa, M., Colombo, H., Saúl, S., Alfonso, A., Combol, A., Castelló, R., and Zanetta, E. Parasitosis intestinales en guarderías comunitarias de Montevideo. *Revista Médica del Uruguay SMU 15*, 1 (Abril 1999), 5–12.

[3] Akaike, H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control 19* (1974), 716–723.

[4] Bates, D., Maechler, M., and Bolker, B. *lme4: Linear mixed-effects models using S4 classes*, 2011. R package version 0.999375-42.

[5] Bhapkar, V. A note on the equivalence of two test criteria for hypotheses in categorical data. *Journal of the American Statistical Association 61* (1966), 228–235.

[6] Blanco, J. *Introducción al Análisis Multivariado.* Instituto de Estadística, FCEA, 2006.

[7] Box, G. A general distribution theory for a class of likelihood criteria. *Biometrika 36* (1949), 317–346.

[8] Cole, T., and Green, P. Smoothing reference centile curves: the lms method and penalized likelihood. *Statistics in Medicine 11* (1992), 1315–1319.

[9] D'Agostino, R. B. Transformation to normality of the null distribution of g_1. *Biometrika 57* (1970), 679–681.

[10] Davis, A. W. Exact distributions of Hotelling's generalized t_0^2 test. *Biometrika 57* (1970), 187–191.

[11] de Onis, M., and Blossner, M. WHO global database on child growth and malnutrition. Tech. rep., World Health Organization, 1997.

[12] Dempster, A. P., Laird, N. M., and Rubin, D. B. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society 39* (1977), 1–38.

[13] Gamboa, M., Navonea, G., Ordenb, G., Torresc, M., Castro, L., and Oyhenartc, E. Socio-environmental conditions, intestinal parasitic infections and nutritional status in children from a suburban neighborhood of La Plata, Argentina. 6.

[14] Hothorn, T., Hornik, K., van de Wiel, M. A., and Zeileis, A. A lego system for conditional inference. *The American Statistician 60*, 3 (2006), 257–263.

[15] Hummel, T. J., and Sligo, J. Empirical comparison of univariate and multivariate analysis of variance procedures,. *Psychological Bulletin 76* (1971), 49–57.

[16] Jurasinski, G., and with contributions from Vroni Retzer. *simba: A Collection of functions for similarity analysis of vegetation data*, 2010. R package version 0.3-2.

[17] Maxwell, A. Comparing the classification of subjects by two independent judges. *British Journal of Psychiatry 116* (1970), 651–655.

[18] Meyer, D., Zeileis, A., and Horni, K. The strucplot framework: Visualizing multi-way contingency tables with vcd. *Journal of Statistical Software 17* (2006), 1–48.

[19] Moustaki, I., and Papageorgiou, I. Latent class models for mixed variables with applications in archaeometry. *Elsevier Computational Statistics & Data Analysis* (Febrero 2004), 17.

[20] Pearson, E. S., and Hartley, H. O. Tables for statisticians. *Biometrika 2* (1972).

[21] PPC. Report of the third global meeting of the partners for parasite control. In *Deworming for Health and Development* (29-30 de Noviembre 2004), World Health Organization, p. 51.

[22] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2010. ISBN 3-900051-07-0.

[23] Rodríguez, D. & Llorca Díaz, M. Estudios longitudinales: Concepto y particularidades. *Revista Española de Salud Pública 78* (2004), 141–148.

[24] Schuurmann, F. J., Krishnaiah, P. R., and Chattopadhyay, A. K. Exact percentage points of the distribution of the trace of a multivariate beta matrix. *Journal of Statistical Computation and Simulation 3* (1975), 331–343.

[25] Schwarz, G. Estimating the dimension of a model. *Annals of Statistics 6* (1978), 461–464.

[26] Stuart, A. A test for homogeneity of the marginal distributions in a two-way classification. *Biometrika 42* (1955), 412–416.

[27] Wall, F. J. The generalized variance ratio or u-statistic, 1967.

Apéndice A

Anexo metodológico

A.1. Z scores

A continuación se detalla el script utilizado para llevar a cabo el cálculo de los puntajes Z correspondientes al peso, la altura y el IMC y algunos de los valores de L, M y S (para cada edad medida en meses) necesarios para calcular dichos indicadores.

	IMC						**Altura**					
	Niños			Niñas			Niños			Niñas		
Edad	L	M	S	L	M	S	L	M	S	L	M	S
61	-0,73	15,26	0,08	-0,88	15,24	0,09	1,00	110,26	0,04	1,00	109,60	0,04
62	-0,76	15,26	0,08	-0,90	15,24	0,09	1,00	110,80	0,04	1,00	110,12	0,04
63	-0,78	15,26	0,08	-0,92	15,24	0,09	1,00	111,33	0,04	1,00	110,64	0,04
64	-0,80	15,26	0,08	-0,94	15,24	0,09	1,00	111,86	0,04	1,00	111,16	0,04
65	-0,83	15,26	0,08	-0,96	15,24	0,09	1,00	112,39	0,04	1,00	111,67	0,04
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
225	-0,88	22,07	0,12	-0,77	21,38	0,14	1,00	176,47	0,04	1,00	163,14	0,04
226	-0,87	22,11	0,12	-0,76	21,40	0,14	1,00	176,49	0,04	1,00	163,15	0,04
227	-0,85	22,15	0,12	-0,75	21,41	0,14	1,00	176,52	0,04	1,00	163,15	0,04
228	-0,84	22,18	0,12	-0,75	21,42	0,14	1,00	176,54	0,04	1,00	163,15	0,04
229	-0,84	22,18	0,12	-0,75	21,42	0,14	1,00	176,54	0,04	1,00	163,15	0,04

```
##########################################
#### Weighted mean  and standard deviation
##########################################

wmean <- function(x,w) { return(rounde(sum(x*w,na.rm=T)/sum(w[!is.na(x)]),digits=2))}
```

```
wsd <- function(x,w) {
   mh <- sum(x*w,na.rm=T)/sum((!is.na(x))*w,na.rm=T)
   sdh<-ifelse(length(x[!is.na(x)])>0,rounde(sqrt(sum(((x-mh)^2)*w,na.rm=T)/
   (sum((!is.na(x))*w,na.rm=T)-1)),digits=2),NA)
   return( sdh )
}

################################################################################################
#### Rounding function - SPlus default rounding function uses the nearest even number rule
################################################################################################

rounde <- function(x,digits=0) {
expo<-10^digits
return(ifelse(abs(x*expo)-floor(abs(x*expo))<0.5,sign(x*expo)*floor(abs(x*expo)),sign(x*expo)*
(floor(abs(x*expo))+1))/expo)
}

################################################################################################
### Function for calculating individual height-for-age z-scores
################################################################################################

calc.zhfa<-function(mat,hfawho2007){
for(i in 1:length(mat$age.mo)) {
if(!is.na(mat$age.mo[i]) & mat$age.mo[i]>=61 & mat$age.mo[i]<229) {
### Interpolated l,m,s values
low.age<-trunc(mat$age.mo[i])
          upp.age<-trunc(mat$age.mo[i]+1)
diff.age<-(mat$age.mo[i]-low.age)
if(diff.age>0) {
l.val<-hfawho2007$l[hfawho2007$age==low.age & hfawho2007$sex==mat$sex[i]]+diff.age*(
hfawho2007$l[hfawho2007$age==upp.age&hfawho2007$sex==mat$sex[i]]-hfawho2007$l[hfawho2007$age==
low.age& hfawho2007$sex==mat$sex[i]])
m.val<-hfawho2007$m[hfawho2007$age==low.age & hfawho2007$sex==mat$sex[i]]+diff.age*(
hfawho2007$m[hfawho2007$age==upp.age&hfawho2007$sex==mat$sex[i]]-hfawho2007$m[hfawho2007$age==
low.age& hfawho2007$sex==mat$sex[i]])
s.val<-hfawho2007$s[hfawho2007$age==low.age & hfawho2007$sex==mat$sex[i]]+diff.age*(
hfawho2007$s[hfawho2007$age==upp.age&hfawho2007$sex==mat$sex[i]]-hfawho2007$s[hfawho2007$age==
low.age& hfawho2007$sex==mat$sex[i]])
} else {
l.val<-hfawho2007$l[hfawho2007$age==low.age & hfawho2007$sex==mat$sex[i]]
m.val<-hfawho2007$m[hfawho2007$age==low.age & hfawho2007$sex==mat$sex[i]]
s.val<-hfawho2007$s[hfawho2007$age==low.age & hfawho2007$sex==mat$sex[i]]
}
mat$zhfa[i]<-(((mat$height[i]/m.val)^l.val)-1)/(s.val*l.val)
} else mat$zhfa[i]<- NA
}
return(mat)
}

################################################################################################
### Function for calculating individual weight-for-age z-scores
################################################################################################

calc.zwei<-function(mat,wfawho2007){
for(i in 1:length(mat$age.mo)) {
if(!is.na(mat$age.mo[i])  & mat$age.mo[i]>=61 & mat$age.mo[i]<121 & mat$oedema[i]!="y"){
### Interpolated l,m,s values
low.age<-trunc(mat$age.mo[i])
          upp.age<-trunc(mat$age.mo[i]+1)
diff.age<-(mat$age.mo[i]-low.age)
if(diff.age>0){
l.val<-wfawho2007$l[wfawho2007$age==low.age & wfawho2007$sex==mat$sex[i]]+diff.age*(
```

```
wfawho2007$l[wfawho2007$age==upp.age&wfawho2007$sex==mat$sex[i]]-wfawho2007$l[wfawho2007$age==
low.age& wfawho2007$sex==mat$sex[i]])
m.val<-wfawho2007$m[wfawho2007$age==low.age&wfawho2007$sex==mat$sex[i]]+diff.age*(
wfawho2007$m[wfawho2007$age==upp.age&wfawho2007$sex==mat$sex[i]]-wfawho2007$m[wfawho2007$age==
low.age& wfawho2007$sex==mat$sex[i]])
s.val<-wfawho2007$s[wfawho2007$age==low.age&wfawho2007$sex==mat$sex[i]]+diff.age*(
wfawho2007$s[wfawho2007$age==upp.age&wfawho2007$sex==mat$sex[i]]-wfawho2007$s[wfawho2007$age==
low.age& wfawho2007$sex==mat$sex[i]])
} else {
l.val<-wfawho2007$l[wfawho2007$age==low.age&wfawho2007$sex==mat$sex[i]]
m.val<-wfawho2007$m[wfawho2007$age==low.age&wfawho2007$sex==mat$sex[i]]
s.val<-wfawho2007$s[wfawho2007$age==low.age&wfawho2007$sex==mat$sex[i]]
}
mat$zwfa[i]<-(((mat$weight[i]/m.val)^l.val)-1)/(s.val*l.val)
if(!is.na(mat$zwfa[i]) & mat$zwfa[i]>3) {
sd3pos<- m.val*((1+l.val*s.val*3)^(1/l.val))
sd23pos<- sd3pos- m.val*((1+l.val*s.val*2)^(1/l.val))
mat$zwfa[i]<- 3+((mat$weight[i]-sd3pos)/sd23pos)
}
if(!is.na(mat$zwfa[i]) & mat$zwfa[i]< (-3)) {
sd3neg<- m.val*((1+l.val*s.val*(-3))**(1/l.val))
sd23neg<- m.val*((1+l.val*s.val*(-2))**(1/l.val))-sd3neg
mat$zwfa[i]<- (-3)+((mat$weight[i]-sd3neg)/sd23neg)
}
} else mat$zwfa[i]<-NA
}
return(mat)
}

##############################################################################################
### Function for calulating individual BMI-for-age z-scores
##############################################################################################

calc.zbmi<-function(mat,bfawho2007){
for(i in 1:length(mat$age.mo)) {
if(!is.na(mat$age.mo[i])&mat$age.mo[i]>=61 & mat$age.mo[i]<229 & mat$oedema[i]!="y") {
### Interpolated l,m,s values
low.age<-trunc(mat$age.mo[i])
          upp.age<-trunc(mat$age.mo[i]+1)
diff.age<-(mat$age.mo[i]-low.age)
if(diff.age>0) {
l.val<-bfawho2007$l[bfawho2007$age==low.age&bfawho2007$sex==mat$sex[i]]+diff.age*(
bfawho2007$l[bfawho2007$age==upp.age&bfawho2007$sex==mat$sex[i]]-bfawho2007$l[bfawho2007$age==
low.age& bfawho2007$sex==mat$sex[i]])
m.val<-bfawho2007$m[bfawho2007$age==low.age&bfawho2007$sex==mat$sex[i]]+diff.age*(
bfawho2007$m[bfawho2007$age==upp.age&bfawho2007$sex==mat$sex[i]]-bfawho2007$m[bfawho2007$age==
low.age& bfawho2007$sex==mat$sex[i]])
s.val<-bfawho2007$s[bfawho2007$age==low.age&bfawho2007$sex==mat$sex[i]]+diff.age*(
bfawho2007$s[bfawho2007$age==upp.age&bfawho2007$sex==mat$sex[i]]-bfawho2007$s[bfawho2007$age==
low.age& bfawho2007$sex==mat$sex[i]])
} else {
l.val<-bfawho2007$l[bfawho2007$age==low.age & bfawho2007$sex==mat$sex[i]]
m.val<-bfawho2007$m[bfawho2007$age==low.age & bfawho2007$sex==mat$sex[i]]
s.val<-bfawho2007$s[bfawho2007$age==low.age & bfawho2007$sex==mat$sex[i]]
}

mat$zbfa[i]<-(((mat$cbmi[i]/m.val)^l.val)-1)/(s.val*l.val)
if(!is.na(mat$zbfa[i]) & mat$zbfa[i]>3) {
sd3pos<- m.val*((1+l.val*s.val*3)^(1/l.val))
sd23pos<- sd3pos- m.val*((1+l.val*s.val*2)^(1/l.val))
mat$zbfa[i]<- 3+((mat$cbmi[i]-sd3pos)/sd23pos)
}
```

```
if(!is.na(mat$zbfa[i]) & mat$zbfa[i]< (-3)) {
sd3neg<- m.val*((1+l.val*s.val*(-3))**(1/l.val))
sd23neg<- m.val*((1+l.val*s.val*(-2))**(1/l.val))-sd3neg
mat$zbfa[i]<- (-3)+((mat$cbmi[i]-sd3neg)/sd23neg)
}

} else mat$zbfa[i]<-NA

}

return(mat)
}
```

A.1.1. Análisis de cluster probabilístico

Aquí se expone el script mediante el cual se obtuvieron los grupos parasitarios. Cabe señalar que, dada la naturaleza iterativa del mismo, fue necesario inicializar el algoritmo desde distintos puntos de arranque. Dichas "semillas" fueron elegidas de manera aleatoria. En última instancia se seleccionaron los resultados que proporcionaran el mayor valor de la verosimilitud.

```
clusterEM2<-function(X,G=3,N=500,pro=TRUE,tau=NULL,tol=1e-3,red=FALSE){
p<-ncol(X); n<-nrow(X)
if(is.null(tau)) tau<-rep(1/G,G)
for (i in 1:G) {
assign(paste('p',i,sep=''),runif(p))
assign(paste('P',i,sep=''),matrix(rep(get(paste('p',i,sep='')),n),n,p,byrow=TRUE))
assign(paste('P',i,sep=''),cbind(get(paste('P',i,sep='')),1-get(paste('P',i,sep=''))))
}
X1<-cbind(X,1-X)
T1<-matrix(NA,N,G);T1[1,]<-tau
P11<-matrix(NA,N,G*p)
cero<-1e-7
iter<-0
maxiter<-N
Cij<-matrix(0,n,G)
er<-1000
phi.n<-phi.v<-NULL
while (iter<maxiter & er>tol){
for (i in 1:G) Cij[,i]<-tau[i]*apply(get(paste('P',i,sep=''))^X1,1,prod)
Cij<-Cij/apply(Cij,1,sum)
tau1<-apply(Cij,2,mean)
den<-apply(Cij,2,sum)
for (i in 1:G) {
phi.v<-c(phi.v,get(paste('P',i,sep=''))[1,1:p])
a<-matrix(apply(Cij[,i]*X1[,1:p],2,sum)/den[i],n,p,byrow=TRUE)
if (any(a>1-cero)) a[a>1-cero]<- 1-cero
b<-1-a
if (any(b>1-cero)) b[b>1-cero]<- 1-cero
assign(paste('P',i,sep=''),cbind(a,b))
phi.n<-c(phi.n,get(paste('P',i,sep=''))[1,1:p])
}
gr<-apply(Cij,1,which.max)
loglik <- sum((X1*log(P1)+log(tau[1]))*ifelse(gr==1,1,0))
for (i in 2:G)loglik <- loglik + sum((X1*log(get(paste('P',i,sep='')))+log(tau[i]))*ifelse(gr==i,1,0))
if (iter>0) er<-abs((loglik-loglik1)/loglik1)
loglik1<-loglik
```

```
if (pro){
par(mfrow=c(2,2),las=1)
  plot(T1[,1],ylim=c(0,1),xlim=c(1,iter),type='l',col='red',xlab='iteracin',ylab='',cex.axis=0.7)
  lines(T1[,2],col='blue')
   lines(T1[,3],col='green')
   plot(0,0,type='n',xlim=c(0,iter),ylim=c(0,1),xlab='iteracin',ylab='',cex.axis=0.7)
   for (i in 1:p) lines(P11[,i],col='red',lty=2)
   plot(0,0,type='n',xlim=c(0,iter),ylim=c(0,1),xlab='iteracin',ylab='',cex.axis=0.7)
   for (i in (p+1):(2*p)) lines(P11[,i],col='blue',lty=2)
   plot(0,0,type='n',xlim=c(0,iter),ylim=c(0,1),xlab='iteracin',ylab='',cex.axis=0.7)
   for (i in (2*p+1):(3*p)) lines(P11[,i],col='green',lty=2)
   print(c(loglik,er,iter))
}
T1[iter+1,]<-tau<-tau1
P11[iter+1,]<-phi.n
iter<-iter+1
phi.n<-phi.v<-NULL
}
bic<- -2*loglik + G*(p+1)*log(n)
aic<- -2*loglik + 2*G*(p+1)
param<-matrix(0,G,p)
for (i in 1:G){
param[i,] <- get(paste('P',i,sep=''))[1,1:p]
}
rownames(param)<-paste('grupo',1:G,sep='')
colnames(param)<-colnames(X1)[1:p]
names(tau)<-paste('grupo',1:G,sep='')
if (!red){
t1<-table(X)
frec<-data.frame(t1,f.esp=0)
for (i in 1:(2^p)) {
f.esp<-0
for (j in 1:G) f.esp<-f.esp+tau[j]*prod(ifelse(frec[i,1:p]==1,param[j,],1-param[j,]))*n
frec[i,p+2]<-f.esp
}
names(frec)<-c(names(X),'f.obs','f.esp')
estadstico<-sum(((frec$f.obs-frec$f.esp)^2)/frec$f.esp)
p.valor<-1-pchisq(sum(((frec$f.obs-frec$f.esp)^2)/frec$f.esp),1)
gl<-1
prueba<-data.frame(estadstico,gl,p.valor)
salida<-list(grupos=G,probs=Cij,loglik=loglik,BIC=bic,AIC=aic,parametros=list(phi=param,tau=tau,
frecuencias=frec,Chisq=prueba,tolerancia=tol,iteraciones=iter)
}else salida<-list(grupos=G,probs=Cij,loglik=loglik,BIC=bic,parametros=list(phi=param,tau=tau))
return(salida)
}
```

A.2. Test de Doornik-Hansen

El siguiente procedimiento pone a prueba la hipótesis de normalidad en un contexto multivariado. La misma se basa en los coeficientes de asimetría y kurtosis de una transformación de los datos.

Las hipótesis nula y alternativa de esta prueba son:

$$H_0)\ Y \sim N_p\left(\mu, \Sigma\right)) \tag{A.1}$$

$$H_1)\ No\ H_0) \tag{A.2}$$

Siendo X la matriz de datos (de n filas por p columnas), $\bar{X}$ la matriz de datos centrados, S la matriz de covarianzas muestral, V una matriz con los inversos de los desvíos standard de cada variable en su diagonal, C la matriz de correlaciones muestrales, Λ una matriz diagonal con los valores propios de C en la diagonal y H la matriz de valores propios de C, se definen los nuevos datos transformados como:

$$R^{'} = H\Lambda^{-1}H^{'}V\bar{X} \tag{A.3}$$

De este modo, cada una de las columnas de R puede considerarse aproximadamente distribuídas como una normal standard.

Sobre estos nuevos datos, se calculan los coeficientes de asimetría y kurtosis de cada variable.

$$B_1^{'} = \left(\sqrt{b}_1, \sqrt{b}_2, \ldots, \sqrt{b}_p\right)^{'} \tag{A.4}$$

$$B_2^{'} = \left(b_1, b_2, \ldots, b_p\right)^{'} \tag{A.5}$$

Estos coeficientes son transformados acorde al procedimiento descrito en D'Agostino [9] de modo de corregir el comportamiento del estadístico cuando la muestra es pequeña. El estadístico en sí es:

$$E_p = Z_1 Z_1^{'} + Z_2 Z_2^{'} \tag{A.6}$$

Siendo Z_1 y Z_2 los vectores de coeficientes transformados de asimetría y kurtosis respectivamente. La distribución del estadístico E_p puede ser adecuadamente aproximada como una χ^2_{2p}.

Aquí se explicita el código utilizado para contrastar, mediante el test de Doornik-Hansen, la hipótesis nula de normalidad (ya sea esta univariada o multivariada).

```
doornik.hansen.test<-function(x){
if (class(x)=='ts') x<-as.numeric(x)
if (class(x)=='numeric'){
b1<-asimetria(x)[1]
b2<-kurtosis(x)[1]
n<-length(x)

beta<-3*((n+1)*(n+3)*(n^2+27*n-70))/((n-2)*(n+5)*(n+7)*(n+9))
w2<-sqrt(2*(beta-1))-1
d<-1/sqrt(log(sqrt(w2)))
y<-b1*sqrt(((w2-1)*(n+1)*(n+3))/(12*(n-2)))
z1<-d*log(y+sqrt(y^2+1))

dk<-(n-3)*(n+1)*(n^2+15*n-4)
a<-((n-2)*(n+5)*(n+7)*(n^2+27*n-70))/(6*dk)
c<-((n-7)*(n+5)*(n+7)*(n^2+2*n-5))/(6*dk)
k<-((n+5)*(n+7)*(n^3+37*n^2+11*n-313))/(12*dk)
alfa<-a+b1^2*c
ji<-2*k*(b2-1-b1^2)
z2<-((ji/(2*alfa))^(1/3)-1+1/(9*alfa))*sqrt(9*alfa)

est<-z1^2+z2^2
gl<-2
pval<-1-pchisq(est,gl)
X<-data.frame(est,gl,pval)
colnames(X)<-c('Estadstico','gl','p-valor')

cat('\n','       ');cat('Test de Doornik-Hansen','\n')
cat(paste('datos:',deparse(substitute(x)),sep=' '),'\n')
print(X)
}
if (class(x)%in%c('matrix','data.frame')){
n<-nrow(x);p<-ncol(x)
media<-apply(x,2,mean)
xc<-sweep(as.matrix(x),2,media)
s<-(1/n)*t(xc)%*%xc
sd<-apply(x,2,sd)
v<-diag(1/sd,nrow=length(sd))
c<-v%*%s%*%v
vp<-eigen(c)
val<-vp$values;h<-vp$vectors
if (any(abs(val)<0.000001)) {
esos<-which(abs(val)<0.000001)
val<-val[-esos]
h<-h[,-esos]
}
lambda<-diag(1/sqrt(val))
r<-h%*%lambda%*%t(h)%*%v%*%t(xc)
b1<-unlist(apply(r,1,asimetria,todo=F))
b2<-unlist(apply(r,1,kurtosis,todo=F))

beta<-3*((n+1)*(n+3)*(n^2+27*n-70))/((n-2)*(n+5)*(n+7)*(n+9))
w2<-sqrt(2*(beta-1))-1
d<-1/sqrt(log(sqrt(w2)))
y<-b1*sqrt(((w2-1)*(n+1)*(n+3))/(12*(n-2)))
z1<-d*log(y+sqrt(y^2+1))

dk<-(n-3)*(n+1)*(n^2+15*n-4)
a<-((n-2)*(n+5)*(n+7)*(n^2+27*n-70))/(6*dk)
c<-((n-7)*(n+5)*(n+7)*(n^2+2*n-5))/(6*dk)
k<-((n+5)*(n+7)*(n^3+37*n^2+11*n-313))/(12*dk)
alfa<-a+b1^2*c
```

```
ji<-2*k*(b2-1-b1^2)
z2<-((ji/(2*alfa))^(1/3)-1+1/(9*alfa))*sqrt(9*alfa)

est<-t(z1)%*%z1+t(z2)%*%z2
gl<-2*p
pval<-1-pchisq(est,gl)
X<-data.frame(est,gl,pval)
colnames(X)<-c('Estadstico','gl','p-valor')

cat('\n','       ');cat('Test de Doornik-Hansen multivariado','\n')
cat(paste('datos:',colnames(x),sep=' '),'\n')
print(X)
invisible(X)
}
}
```

Apéndice B

Anexo estadístico

B.1. Análisis de cluster

Establecidas las variables y los objetos a clasificar, es necesario definir una medida de distancia entre ellos que cuantifique el grado de similaridad (o disimilaridad) entre cada par de objetos. Esta distancia también será aplicada para medir la proximidad entre grupos y entre objetos y grupos. Dependiendo del tipo de variables a analizar existen varios tipos de distancia. Las distancias que se detallan a continuación pueden ser tenidas en cuenta para variables cuantitativas:

- $d_{ij}^2 = \sum_{k=1}^{p}(x_{ik} - x_{jk})^2$ distancia euclideana.
- $d_{ij} = \left(\sum_{k=1}^{p}|x_{ik} - x_{jk}|^h\right)^{1/p}$ distancia de Minkowski.
- $d_{ij}^2 = (x_i - x_j)' S^{-1}(x_i - x_j)$ distancia de Mahalanobis.

En caso de que las variables sean binarias el procedimiento a seguir es algo diferente. Para cada par de individuos deberá construirse la siguiente tabla:

	ind_i	
ind_j	a	b
	c	d

Donde a es el número de atributos que los individuos comparten y d es la cantidad de atributos ausentes en ambos individuos. Por su parte, b y c son aquellos atributos que se encuentran presentes en un individuo pero no en el otro. En función de estos valores pueden definirse una gran cantidad de índices de semejanza, algunos de los más utilizados son[15]:

- $s_{ij} = \frac{d}{b+c+d}$ coeficiente de Jaccard
- $s_{ij} = \frac{a+d}{a+b+c+d}$ coeficiente de acuerdo simple

Los métodos de clasificación pueden ser jerárquicos o no jerárquicos. Los métodos jerárquicos pueden ser de carácter agregativo o divisivo. En los primeros la agrupación se lleva a cabo de forma tal que objetos que estén cerca uno del otro (con respecto de cierta distancia) conformarán un grupo. Este proceso se repite sucesivamente hasta que todos los individuos conformen un único grupo. En el caso de los métodos divisivos, el procedimiento es el inverso. Mediante una sucesión de particiones los grupos se van subdividiendo hasta llegar a la instancia en que cada individuo conforma un grupo por sí mismo. A diferencia de los anteriores, los métodos no jerárquicos permiten la reasignación de individuos a otros grupos conforme el algoritmo de agrupación va avanzando.

En este estudio, los algoritmos utilizados fueron de carácter agregativo. A continuación se detallan algunos:

- Vecino mas cercano. En este algoritmo el criterio utilizado para medir la distancia entre grupos o entre individuos y grupos, es el siguiente:

$$d_{A,B} = \text{mín}\, d_{ij} / x_i \in A, x_j \in B \tag{B.1}$$

  Es decir, de todas las distancias que involucren un individuo del grupo A y uno del B, se selecciona la menor de ellas. Una vez re-calculadas todas las distancias entre elementos (grupos y/o individuos) se fusionan aquellos que disten lo menos posible entre sí. Este procedimiento se repite hasta que todos los individuos conformen un único grupo.

- Vecino mas lejano. El criterio utilizado en este algoritmo para medir la distancia entre grupos o entre individuos y grupos, es el siguiente:

$$d_{A,B} = \text{máx}\, d_{ij} / x_i \in A, x_j \in B \tag{B.2}$$

Así, de todas las distancias que involucren un individuo del grupo A y uno del B, se selecciona la mayor de ellas. Una vez re-calculadas todas las disntancias entre elementos (grupos y/o individuos) se fusionan aquellos que disten lo menos posible entre sí. Este procedimiento se repite hasta que todos los individuos conformen un único grupo.

- Método de Ward. Este método se caracteriza por emplear las distancias "dentro" y "entre" los grupos. El método de Ward combina los grupos A y B de tal forma que el incremento en la suma de cuadrados dentro del nuevo grupo con respecto a los anteriores sea mínimo.

$$\Delta_{sc} = SCD_{AB} - (SCD_A - SCD_B) \tag{B.3}$$

Donde SCD representa la suma de cuadrados dentro de cada uno de los grupos. Al hablar de sumas de cuadrados se refiere a:

$$SCD_A = \sum_{j=1}^{p} \sum_{i,j \in A} (x_{ij} - \bar{x}_j)^2 \tag{B.4}$$

Este proceso se repite hasta que todos los objetos conformen un solo grupo.

B.1.1. Determinación del número de grupos

Pese a que en la literatura existen numerosos indicadores que permiten escoger una cierta cantidad de grupos, a continuación se exponen los utilizados en este trabajo.

- R^2: Este representa la relación entre la variación explicada por la estructura de los K grupos y la variación total.

$$R^2 = 1 - \frac{\sum_{k=1}^{K} \sum_{i=1}^{n_k} \sum_{j=1}^{p} (x_{ij(k)} - \bar{x}_{j(k)})^2}{\sum_{i=1}^{n_k} \sum_{j=1}^{p} (x_{ij} - \bar{x}_j)^2} \tag{B.5}$$

Cuando se tiene N grupos (cada individuo es un grupo), $R^2 = 1$. Por otro lado, cuando todos los individuos conforman un único grupo $R^2 = 0$. El número de grupos K se selecciona cuando el incremento en el R^2 al pasar de K a $K+1$ deja de ser "significativo".

- *Pseudo F*

  Este indicador se construye pensando en los datos como si se tratase de poner a prueba la significación de la variable de agrupación bajo un diseño de análisis multivariado de la varianza a una vía. De esta forma el indicador adopta la siguiente expresión:

$$pseudoF = \frac{\frac{tr(SCE)}{K-1}}{\frac{tr(SCD)}{N-K}} \tag{B.6}$$

  Si se piensa al indicador como una función de la cantidad de grupos, se elige el número K de grupos en el cual el indicador presenta un máximo relativo.

B.2. Test Box-M

El test que se describe a continuación es útil para poner a prueba el supuesto de homogeneidad de matrices varianzas y covarianzas. El procedimiento es muy flexible en el sentido en puede ser usado independientemente de que el modelo sea balanceado o que el diseño sea a una o dos vías.

La hipótesis nula y alternativa de esta prueba son:

$$H_0)\ \Sigma_1 = \Sigma_2 = \ldots \Sigma_k \tag{B.7}$$

$$H_1)\ \Sigma_i \neq \Sigma_j \text{ para algún } \ i \neq j \tag{B.8}$$

En la construcción del estadístico se necesitarán las matrices de covarianzas correspondientes a cada uno de los grupos (S_i) y los grados de libertad de cada uno de ellos (ν_i).

Al disponer de estos elementos el estadístico se formula de la siguiente manera:

$$M = \frac{\prod_{i=1}^{k} |S_i|^{\frac{\nu_i}{2}}}{|S_{pool}|^{\sum_{i=1}^{k} \frac{\nu_i}{2}}} \tag{B.9}$$

Es importante notar que la desigualdad $\nu_i > p$ debe satisfacerse ya que en caso contrario los determinantes valdrían cero y por ende, el estadístico también. Donde S_{pool} no es más que la matriz de covarianzas combinada, es decir:

$$S_{pool} = \frac{\sum_{i=1}^{k} \nu_i S_i}{\sum_{i=1}^{k} \nu_i} \tag{B.10}$$

En su trabajo original de 1949, Box [7] ofreció una aproximación a la distribución del estadístico correspondiente a la ecuación 3.50. Para que la distribución del mismo se aproxime a una χ^2, la transformación debe ser la siguiente:

$$2\,(1 - c_1)\,log\,(M) \tag{B.11}$$

Donde:

$$c_1 = \left(\sum_{i=1}^{k} \frac{1}{\nu_i} - 1\sum_{i=1}^{k} \nu_i\right)\left(\frac{(2p^2 + 3p - 1)}{6\,(p+1)\,(k-1)}\right) \tag{B.12}$$

Los grados de libertad del mismo serán:

$$gl = \frac{1}{2} p\,(p+1)\,(k-1) \tag{B.13}$$

B.3. Criterios de selección de modelos

En el caso de que varios modelos proporcionen un ajuste adecuado para el mismo problema, es natural que surja la interrogante: ¿cuál de ellos proporciona el mejor ajuste? En este apartado se definen los dos criterios de selección de modelos mas usados y difundidos en la literatura especialiada. Sea $L_n(\theta)$ la log-verosimilitud de un cierto modelo especificado a través del vector de parámetros θ y basado en una muestra de tamaño n. Siendo p_0 la cantidad correcta de parámetros se definen dos situaciones:

- Modelos con $p > p_0$ estarán sobre-parametrizados.
- Mientras que, en el caso de que $p < p_0$, el modelo estará mal especificado.

Los criterios de selección de modelos mas comúnmente encontrados en la práctica son:

- AIC: $-2log(L_n(\theta)) + 2p$

  El criterio de información Akaike es una medida de la relativa bondad de ajuste de un modelo estadístico. Fue desarrollado por Hirotsugu Akaike [3].

- BIC: $-2log(L_n(\theta)) + log(n)p$

  El criterio de información bayesiana fue desarrollado por Gideon E. Schwartz [24] quien adoptó una postura bayesiana en su formulación.

La diferencia entre estos indicadores radica en la forma en que penalizan la inclusión de parámetros en la formulación del modelo. En ambos casos, será seleccionado el modelo que adopte el menor valor del indicador.

B.4. Función Discriminante

El rechazo de la hipótesis nula implica que $\mu_{Aj} \neq \mu_{Bj}$ para al menos un j. Sin embargo, nada garantiza que las pruebas univariadas lleguen a esta conclusión. No obstante se puede considerar una combinación lineal de las p variables, de la forma $z = a^{'}y$, tal que para cierto vector de coeficientes a, la siguiente prueba será significativa;

$$t(a) = \frac{\bar{z}_A - \bar{z}_B}{\sqrt{(1/n_A + 1/n_B)S_z^2}} \tag{B.14}$$

Siendo la hipótesis a testear:

$$H_0)\ \mu_{Z_A} = \mu_{Z_B} \tag{B.15}$$

ó lo que es equivalente;

$$H_0)\ \ a^{'}\mu_A = a^{'}\mu_B \tag{B.16}$$

En el caso que $a = S_{pool}^{-1}(\bar{y}_A - \bar{y}_B)$, $a^{'}y$ es llamada "función discriminante". Es importante mencionar que al elegir el vector de coeficientes a de esta manera, se está proyectando sobre la dirección que maximiza la distancia estandarizada entre los vectores de medias. La utilidad de la función discriminante radica en que al examinar los coeficientes a_j podemos tener una idea sobre qué variables contribuyen en mayor medida a la significación del test T^2 llevado a cabo en primer instancia.

B.5. Homogeneidad marginal

B.5.1. Procedimientos post hoc

En aquellas situaciones en las que el estadístico W_0 alcance un valor lo suficientemente grande como para rechazar la hipótesis de homogeneidad marginal, la siguiente pregunta puede ser de particular interés, ¿cuál o cuáles de las I categorías son significativamente diferentes? Para responder esta pregunta basta con construir I nuevas variables indicadoras que registren la presencia o ausencia de la i-ésima categoría. De este modo podrán construirse I tablas de dos por dos en las cuales se podrá hacer hincapié en la diferencia (o no) de las proporciones marginales de la categoría i. Cabe señalar que al aplicar el estadístico W_0 (en la versión de Stuart-Maxwell) sobre estas tablas reducidas, el mismo es equivalente al estadístico Q de la prueba de McNemar.

B.5.2. Homogeneidad marginal para variables ordinales

Al equipo del Instituto de Higiene le pareció pertinente considerar al análisis de ciertos indicadores ordinales construidos a partir de la discretización de los Z scores. A modo de ejemplo, el Z score correspondiente a la talla se discretizó en tres categorías (normal, retraso y retraso severo). A través de la tabla de contingencia que cruza las dos etapas (indicador talla inicial vs indicador talla final) se pretendió analizar el cambio antropomeétrico de forma cualitativa.

En el caso de que las categorías de la variable a testear tengan un ordenamiento natural, la prueba que se describe a continuación puede resultar de utilidad en los ocasiones en las que el investigador se pregunte si en una segunda instancia las observaciones tienen a registrarse en valores mayores (o menores) que en la primera. Un modelo que permite analizar este problema es el siguiente logit de odds proporcionales:

$$logit[P(Y_{i1} \leq j)] = \alpha_{ij} + \beta \tag{B.17}$$

$$logit[P(Y_{i2} \leq j)] = \alpha_{ij} \tag{B.18}$$

El parámetro de interés en esta configuración resulta ser β, ya que el mismo se interpreta en el sentido de que, para los dos momentos de cada individuo los odds de que en la primer instancia registre un valor menor o igual a j son e^{β} veces los odds en la segunda instancia. De esta manera, poner a prueba la hipótesis de homogeneidad marginal equivale a llevar a cabo la siguiente prueba:

$$H_0)\ \beta = 0 \tag{B.19}$$

$$H_1)\ \beta \neq 0 \tag{B.20}$$

Un estimador de β en el modelo descrito anteriormente podría ser el siguiente:

$$\hat{\beta} = log\left(\frac{\sum\sum_{i<j}(j-i)n_{ij}}{\sum\sum_{i>j}(i-j)n_{ij}}\right) \tag{B.21}$$

$$s_{\hat{\beta}} = \sqrt{\frac{\sum\sum_{i<j}(j-i)^2 n_{ij} \;+\; \sum\sum_{i>j}(i-j)^2 n_{ij}}{\left[\sum\sum_{i<j}(j-i) n_{ij}\right]^2 \;+\; \left[\sum\sum_{i>j}(i-j) n_{ij}\right]^2}} \tag{B.22}$$

Una vez obtenidas dichas estimaciones, el cociente $\hat{\beta}/s_{\hat{\beta}}$ se distribuye aproximadamente como una variable aleatoria normal estándar.