16. Gestión y análisis de datos *

Gestión de datos: estrategias y problemas en la recolección, procesamiento, documentación, y resumen de datos para un estudio epidemiológico.

1. Gestión de Datos

1.1 Introducción a la Gestión de Datos

La Gestión de Datos cae bajo el capítulo de gestión de proyectos. La mayor parte de los investigadores no están preparados para la gestión de un proyecto, dado que se tiende a ponerle poco énfasis en el tema en los programas de entrenamiento. Un proyecto epidemiológico se asemeja a llevar a cabo un proyecto de negocios, con una diferencia fundamental, el proyecto tiene una vida de duración predeterminada. Esta diferencia influirá en muchos aspectos de su gestión. Algunas áreas de la gestión que se ven afectadas son la contratación, el despido, la evaluación, la organización, la productividad, el estado de ánimo, la comunicación, la ética, el presupuesto, y la finalización del proyecto. Aunque el desarrollo de una propuesta de investigación plantea muchos desafíos de gestión, si la propuesta es aprobada y se le adjudica la financiación, los logros del proyecto dependen más de la gestión del mismo que de cualquier otro factor.

Un problema particular para el investigador y el personal, cuando no tienen entrenamiento o experiencia específica, es que no logran apreciar y prepararse para lo que implica y exige la producción en masa.

1.2 El Sistema de Gestión de Datos

El sistema de gestión de datos es el conjunto de procedimientos y personas por medio de los cuales se procesa la información. Involucra la recolección, manipulación, almacenamiento, y recuperación de información. Talvez la herramienta más visible es la computadora; sin embargo, es meramente una de tantas herramientas necesarias. Otras "herramientas" son los instrumentos y los formularios de recolección de datos, el protocolo de gestión de datos, los mecanismos de control de calidad, documentación, instalaciones de almacenamiento tanto para el papel como los medios electrónicos, y los mecanismos de recuperación. El objetivo del sistema de gestión de datos es el de asegurar: a) datos de alta calidad, i.e., asegurar que la variabilidad en los datos proviene del fenómeno en estudio y no del proceso de recolección de datos, y b) un análisis e interpretación de datos precisos, apropiados y defendibles.

---

*La versión original de este capítulo fue escrita por H. Michael Arrighi, Ph.D

1.3 Objetivos específicos de la Gestión de Datos

Los objetivos específicos de la gestión de datos son:

1.3.1 Obtener los datos y prepararlos para el análisis

El sistema de gestión de datos incluye la supervisión del flujo de datos desde los sujetos de investigación a los analistas de datos. Antes de poder analizarlos, los datos deben ser recogidos, revisados, codificados, computarizados, verificados, confirmados y convertidos a formularios adecuados para llevar a cabo el análisis. El proceso debe ser adecuadamente documentado para fundamentar el análisis e interpretación.

1.3.2 Mantener el control de calidad y la seguridad de los datos

Las amenazas con respecto a la calidad de los datos surgen en todos los puntos en que se obtienen y/o modifican datos. El valor de la investigación se verá muy afectado por el control de calidad, pero lograr y mantener la calidad requiere actividades que a menudo son banales y difíciles de motivar. El control de calidad incluye:

- Prevenir y detectar errores en los datos a través de procedimientos escritos, entrenamiento, procedimientos de verificación, y evitando complejidades innecesarias
- Evitar o eliminar las inconsistencias, errores, y datos faltantes a través de la revisión de los formularios de recolección de datos (en forma ideal cuando el acceso a las fuentes de los datos aún está disponible para permitir resolver las dudas) y los conjuntos de datos.
- Evaluar la calidad de los datos a través de los apuntes de los entrevistadores, codificadores, editores de datos, a través del interrogatorio de los sujetos, y a través de revisiones o repeticiones de la recolección de datos para sub-muestras
- "Sentir" los datos, evitar interpretaciones equivocadas y descuidos importantes

Los temas de seguridad incluyen: (1) lo legal, (2) la seguridad de la información, (3) la protección de las fuentes externas, (4) la protección de las fuentes internas. Mientras que el abuso es más llamativo, los problemas accidentales son más comunes. Las medidas de prevención típicas son la separación o aislamiento de la información que identifica los sujetos investigados (para proteger la confidencialidad), la redundancia, los respaldos (para protegerse contra malfuncionamiento humano y de las máquinas). La pérdida de datos importantes debido a fallas en tener una copia de respaldo segura puede ser considerada negligencia. Lamentablemente, puede existir una relación inversa entre la seguridad y la accesibilidad/utilidad de los datos.

1.3.3 Permitir las solicitudes de información, revisión, reconstrucción, y archivo

En cualquier momento durante y hasta después de que se completa el proyecto puede surgir la solicitud de los instrumentos y/o de los datos o preguntas sobre los mismos. La agencia

financiadora requerirá un informe final. Otros investigadores o grupos interesados (p.ej. empresas cuyos productos están implicados como amenazas a la salud) pueden solicitar una copia del conjunto de los datos para realizar sus propios análisis. Raramente se llevará a cabo una investigación por lo sorprendente de los hallazgos, el involucramiento de personas con una importante inversión en sus implicancias, o sospechas o acusaciones con respecto al estudio. Por ejemplo, Herbert Needleman, un investigador pionero de los efectos de la exposición al plomo en los niños sobre las funciones cognitivas, hizo que sus datos y sus resultados fueran auditados por una comisión científica (que incluía un profesor de UNC). Proctor y Gamble, Inc. Le hicieron un juicio al CDC para que tuviera que entregar los datos de sus estudios caso-control sobre el síndrome de choque tóxico y tampones.

Continúan en aumento las preocupaciones por la mala conducta científica y el fraude, y los investigadores tienen la responsabilidad de mantener la documentación necesaria para refutar acusaciones de ese tipo si llegaran a presentarse. En forma cada vez más frecuente, las revistas exigen que los datos (y toda la documentación pertinente que los apoya) sea mantenida durante varios años después de la publicación. En un nivel más práctico, surgirán numerosas preguntas a lo largo del análisis de datos, y el sistema de gestión de datos del proyecto tiene que poder responder en forma precisa y oportuna.

Un principio importante en la gestión de datos, en todos los niveles y etapas, es la contabilidad completa de los datos. Así cuando se desarrolla la actividad de recolección de datos, debe existir un registro detallado del número de sujetos (si se conoce) en el universo del cual se reclutan los sujetos y una tabulación completa dentro de un conjunto de categorías mutuamente excluyentes. Las categorías típicas son: no elegibles según la razón por la cual no pueden ser seleccionados (p.ej., edad fuera del rango, condiciones médicas), no-participantes según las razones para no participar (p.ej., no tiene número de teléfono, teléfono desconectado, fuera de la ciudad, se niega), participantes cuyos datos son excluidos (p.ej., demasiados datos faltantes, el entrevistador tiene dudas de que el participante esté diciendo la verdad), etc.

Un registro de auditoría es un mecanismo esencial para identificar los cambios que son realizados a los datos en cada paso. Este registro debe documentar qué cambios se realizaron, quién los realizó, dónde, cuándo y cómo fueron realizados. Los registros de auditoría son importantes para responder a o recuperarse de: (1) enfrentamientos legales, (2) problemas de procedimientos, (3) problemas menores, y (4) desastres

Señalemos que los objetivos anteriores se aplican tanto a los sistemas manuales como los computarizados.

1.3.4 Temas particulares de los estudios ciegos

La epidemia de VIH ha llevado a que se realicen más encuestas serológicas ciegas para determinar la prevalencia de la infección por VIH en distintos ambientes, subgrupos, y áreas geográficas. Para evitar el sesgo de no-respuesta, una preocupación especial en los estudios sobre VIH dada la baja prevalencia del virus en la mayor parte de las poblaciones y el miedo y estigma asociados con la infección por VIH y sus factores de riesgo, se han desarrollado métodos para realizar estudios ciegos (no vinculados). Dichos estudios usan sangre sobrante extraída originalmente

con otro objetivo (p.ej., pruebas médicas) y que son analizadas de manera que es imposible obtener la identificación de los individuos involucrados en el estudio. En circunstancias particulares, estos estudios no requieren el consentimiento informado, de manera que son libres del sesgo de no-respuesta.

Se requiere un cuidado especial para diseñar un sistema de gestión de datos que pueda prevenir la posibilidad de vincular los datos con los sujetos individuales. Por ejemplo, los procedimientos habituales de recolección de datos como el uso de números de identificación secuenciales, inclusión de las fechas exactas en todos los formularios, y el registro de información adicional para aclarar los datos que son dudosos pueden comprometer el anonimato. De hecho, los estudios de datos no-vinculados, generan un conflicto básico entre la necesidad de evitar la vinculación y los objetivos de gestión de los datos principales, como son el monitoreo y control de calidad, que requieren la posibilidad de volver atrás y verificar información.

1.4 Los componentes de la Gestión de Datos

1.4.1 Gestión

Los conceptos de gestión se pueden aplicar tanto a la gestión de datos como a la gestión de proyectos. Los temas de gestión son componentes críticos de los sistemas de gestión de datos. Los datos son simplemente objetos que son manipulados por el sistema de gestión de datos. Los datos no tendrán importancia si no se le presta una atención adecuada al proceso.

El equipo de investigación es el responsable último del resultado del proyecto. Aún en los proyectos importantes en que existe un gerente del proyecto y un gerente de datos, el equipo de investigación es el comité de dirección del proyecto. Se requieren destrezas de gestión para evaluar a los administradores y asegurarse que están cumpliendo adecuadamente su trabajo, además de mantener el proyecto según el calendario previsto. Aún en un proyecto relativamente pequeño, puede ser necesario que los investigadores tengan que esforzarse para cumplir con el rol de administradores, ya que muchas de las cualidades que hacen a un buen investigador son opuestas a las que hacen a un buen administrador:

Investigador	Administrador
Soluciones óptimas	Soluciones pragmáticas
Soluciones precisas	Soluciones prácticas
Trabaja con cosas	Trabaja con personas
Orientado a los procesos	Orientado a los resultados
Éxito individual	Éxito grupal

Un buen investigador requiere creatividad y puede ser considerado "quisquilloso", i.e. una persona que constantemente está cambiando las cosas basándose en nuevas ideas. Un buen administrador también es creativo pero es menos quisquilloso. Cambios constantes en la situación de administración o gestión crean confusión y falta de consistencia, que en última instancia resulta en datos de mala calidad. Algunos de los elementos de gestión claves que afectan directamente el sistema de gestión de datos son:

1.4.1.1 Comunicación bidireccional

Cualquier persona que trabaja en el proyecto puede hacer un comentario u observación o una contribución valiosa. Estas contribuciones deben ser respetadas. El hecho de escuchar es un aspecto importante para enterarse de lo que realmente está ocurriendo y tratar de encontrar soluciones. La persona que lleva a cabo la tarea a menudo es la que mejor conoce los detalles y lo que es efectivo mejor que cualquier otro. Las personas tienen distinto grado de facilidad para expresar sus ideas y preocupaciones, y son importantes las oportunidades para las discusiones informales frente a frente (p.ej. tomando un café).

1.4.1.2 Consistencia

La consistencia es esencial para la implementación del protocolo, en el proceso de recolección de datos, y con respecto a las decisiones tomadas durante el proyecto. La falta de consistencia puede resultar de la toma de distintas decisiones por parte de los investigadores principales, una falta de comunicación en la transmisión de problemas y soluciones, y toma de distintas decisiones sobre el mismo tema en diferentes momentos. Surgirán innumerables problemas menores (y muchos problemas mayores) durante el proyecto. Una perspectiva útil para resolver estos temas es preguntarse como lo apreciaría un investigador externo. A menudo las decisiones deben tomarse en forma rápida y deben ser registradas de alguna manera que permita que puedan ser consultadas sin dificultades cuando surja la misma pregunta o una similar en el futuro.

Una incapacidad para implementar el protocolo de estudio en forma consistente puede resultar en un sesgo de selección o un sesgo de información. El sesgo de información en las variables de confusión puede comprometer la capacidad de corregir el fenómeno de confusión en el análisis. Además, cuando se presentan los métodos y los resultados a otros (una parte fundamental de la investigación académica) y surgen las inevitables preguntas, da un poco de vergüenza tener que describir y explicar métodos inconsistentes.

1.4.1.3 Líneas de autoridad y responsabilidad

La autoridad y la responsabilidad deben ser claramente definidas y las personas designadas accesibles para el personal. La accesibilidad es a menudo un problema en los proyectos de investigación académicos en que la mayor parte del personal s están a tiempo parcial, y el personal de tiempo completo tiene otros compromisos que resultan en una falta de accesibilidad para el personal. Entre otras cosas generalmente es deseable designar una persona que sea responsable de autorizar todos los cambios a los datos computarizados.

1.4.1.4 Flexibilidad

El sistema de manejo de datos debe ser flexible para responder a los cambios en el protocolo, instrumentos de encuesta y los cambios de personal. Cuanto más tiempo lleva el proyecto, más susceptible es a los cambios. Todo proyecto sufrirá algunas modificaciones. Así el sistema de gestión de datos debe ser suficientemente flexible para permitir modificaciones sencillas.

1.4.1.5 Simplicidad

Mantén el sistema de gestión de datos lo más simple posible dentro de las posibilidades del (potencial) personal. La simplicidad disminuye los errores disminuyendo la dependencia sobre la "personal clave" y permitiendo que el sistema sea más fácil de aprender e implementar.

Las computadoras son herramientas maravillosas en la gestión de datos, pero es fácil complicar las cosas usándolas. El uso de programas no amigables para el usuario o paquetes no comunes genera complejidad. Los sistemas computarizados en realidad aumentan el costo y el apoyo técnico de los sistemas. Los beneficios que proveen están en el área de una mayor eficiencia y (con suerte) una disminución en los errores. Un proyecto pequeño puede beneficiarse de un sistema predominantemente manual cuando está correctamente diseñado e implementado.

1.4.2 Integración

Es aconsejable integrar el sistema de gestión de datos en todo el proceso de estudio desde la etapa de la idea y la propuesta hasta el trabajo impreso final, el almacenamiento de la información y la destrucción planificada. Obviamente, se presta cierta atención a la gestión de datos en la etapa de propuesta durante el proceso de presupuesto y recursos humanos. Se necesita más atención; se debe pensar en un flujo general del sistema. Esto dará una evaluación preliminar de la demanda de recursos y de la factibilidad.

1.4.3 Estandarización

La estandarización se extiende no sólo a los instrumentos sino también a los procedimientos para revisión de las fichas, los mecanismos de ingreso de datos, la documentación y cualquier otra faceta. Es imprescindible para obtener información de calidad.

1.4.4 Prueba Piloto

Se realiza habitualmente la prueba piloto con los instrumentos de la encuesta. Raramente se hace una prueba piloto con los elementos fundamentales del sistema de gestión de datos. Utiliza la prueba piloto de los instrumentos de la encuesta como una oportunidad para hacer una prueba piloto de ciertos aspectos del sistema de gestión de datos, p.ej., coordinación de entrevistadores, retorno de llamadas, coordinación con otras fuentes (identificación del participante) etc. Un aspecto clave de todo esto es tratar de hacerlo lo más parecido a la realidad posible para evitar el síndrome de la "prueba piloto" y la falta de seriedad por parte del personal.

El sistema de gestión de datos puede ser ensayado cuando las versiones preliminares de los instrumentos de la encuesta están siendo evaluadas y durante la evaluación de los métodos de laboratorio. Si el proyecto es suficientemente grande, se puede realizar una prueba piloto de todo el sistema utilizando los primeros participantes (5, 10 o 20). Luego el proyecto se frena brevemente para su revisión y modificación antes de completar la implementación. Los grandes proyectos usan una cohorte de "vanguardia" que atraviesa todos los aspectos del estudio con suficiente anticipación como para permitir el ajuste de instrumentos y procedimientos.

1.4.5 Control de Calidad y Garantía de Calidad

1.4.5.1 Redundancia

Un procedimiento bien conocido y utilizado de control de calidad es una duplicación de la recolección de algunos datos en la encuesta. Esto se aplica de igual manera a los instrumentos de la encuesta, los procedimientos de laboratorio, y el sistema de flujo de datos. La duplicación puede realizarse en serie o en paralelo.

1.4.5.1.1 En paralelo

El control en paralelo significa la evaluación simultánea de dos ítems de recolección de datos. Cuando es un dato de laboratorio, significa la presentación ciega de dos ítem idénticos para su evaluación. Con un instrumento de encuesta, se trata de la repetición de una pregunta, tal vez planteada en un formato levemente distinto.

1.4.5.1.2 En Series

El control en series es la repetición de dos ítems en dos momentos diferentes. En el caso del laboratorio, es la presentación ciega de dos ítems idénticos en distintos momentos. Con un instrumento de encuesta, es la repetición de todo o una parte del instrumento de encuesta en un momento diferente. Esto puede involucrar la citación de nuevo de una muestra del grupo original para contestar un cuestionario breve de verificación que interroga sobre ítems similares a los del cuestionario original. Se comparan las respuestas de los dos cuestionarios, se identifican las no coincidencias y se vuelven a ingresar. El ingreso por duplicado de los datos (también llamado verificación clave) aunque es habitual no es automática, de manera que generalmente debe ser específicamente solicitada y presupuestada.

1.4.5.2 Introducción de errores

Una técnica útil es la de introducir errores en el sistema de gestión de datos para evaluar los mecanismos de detección y la consistencia del manejo de los errores. Esto se puede realizar ingresando datos erróneos o identificando un problema particular y siguiéndolo a través del sistema de gestión de datos.

1.4.6 Disminuye el número de puntos de ingreso de datos

Cada participante debe ingresar en el estudio de la misma manera o cada sujeto debe tener la misma oportunidad para poder ingresar al estudio. No se deben usar distintos protocolos para ingresar distintos sujetos de la misma categoría (p.ej. casos o controles, o los expuestos y los no expuestos). Esto puede ser un gran desafío cuando el estudio es multicéntrico.

Toma un enfoque planificado a todas las fuentes de datos, incluyendo los registros, formularios de seguimiento y sistemas de citas. La retrospección puede revelar la utilidad de fuentes de datos que no estaban destinadas a ser incluidas originalmente en el análisis. Las consideraciones en la planificación incluyen el diseño del procedimiento de recolección, registro, codificación,

informatización, verificación, garantía de seguridad, y documentación. Trata de limitar las situaciones en que se realizan cambios directamente en la base de datos, sin un registro de auditoria. Sin este registro de auditoria puede ser imposible reconstruir los datos si se realiza un cambio equivocado o aún verificar si se ha realizado un cambio.

1.4.7 Controla

Controla la gestión de datos para asegurar su correcta implementación. Por ejemplo, cuando se está llevando a cabo una actividad de recolección de datos debería realizarse una revisión frecuente y regular del número de sujetos participantes, las razones de la no-participación, los problemas que se han encontrado, etc. Los formularios de recolección de datos (o una muestra de ellos si el volumen es importante) deben ser examinados precozmente para identificar problemas (por ejemplo, un exceso de datos que faltan, ítems no comprendidos) para los cuales se pueden tomar acciones correctivas. Es importante tener un sistema manual o informatizado para controlar los formularios de datos.

Los elementos esenciales para identificar son:

1. Adherencia al protocolo de estudio (para asegurar el mantenimiento de los objetivos de estudio);
2. Consistencia en la implementación del protocolo;
3. Debilidades (y fortalezas) del sistema de gestión de datos;
4. Respuesta a los cambios, los problemas, las crisis - ¿qué tan bien detecta el sistema de gestión de datos los cambios, problemas y crisis y cómo responde a ellos? Este control puede llevarse a cabo utilizando datos erróneos o haciendo un seguimiento de la fecha y los ítems identificados como problema, la fecha de su identificación y la fecha de su corrección.

1.4.8 Documentación

Documentar es un desafío especial por su falta de atractivo y por el hecho de que en cualquier momento particular del estudio (antes de su finalización) las prioridades urgentes compiten para hacer que la documentación sea una actividad muy difícil de mantener. Sin embargo es una actividad absolutamente imprescindible y no puede ser siempre reconstruida después de que ocurren los hechos. Hay que presupuestar el tiempo y el personal para la documentación de los eventos, las decisiones, los cambios, los problemas y las soluciones. Revisa la documentación a medida que es producida para mostrar la importancia que le asignas y para asegurarte que se está produciendo de la manera que la necesitas.

Documenta las reuniones de los investigadores y las reuniones de los comités poniendo todos los temas en la agenda (debería alcanzar con una o dos frases), seguidos por la decisión o acción tomadas (abierto, cerrado, resuelto y el resumen del mismo). El relato de toda la discusión es interesante pero tiende a ser demasiado largo – trata de ser breve y anotar los puntos claves. La obligación de tener las notas publicadas en uno o dos días después de la reunión obliga a que sean breves y obliga a que sean escritas antes de que la memoria las afecte. Otra técnica es mantener un

diario contemporáneo o un documento en procesador de texto con notas fechadas sobre las cosas que deben ser incluidas en los informes de progreso.

La documentación del proyecto debe incluir:

- Un breve relato de los objetivos y los métodos del estudio
- Una cronología detallada de los eventos y actividades durante el trayecto del proyecto, mostrando las fechas de comienzo y finalización y el número de sujetos para cada actividad de recolección de datos.
- Para cada actividad de recolección de datos, un registro de los procedimientos utilizados en detalle y una contabilidad del número de sujetos incluidos, el número seleccionado para la recolección de datos, y el resultado final para todos los sujetos (por categoría, p.ej., no se pudo contactar, se niega). Este material debe tener referencias cruzadas con las fuentes originales (p.ej. en la computadora) para permitir la verificación cuando sea necesaria.
- Un compendio de todos los instrumentos de recolección de datos, incluyendo la documentación de las fuentes para las preguntas del cuestionario obtenidas de instrumentos pre-existentes. Deben incluirse los resultados de pruebas previas y análisis de validación o referencias cruzadas a ellos.
- Listas y descripciones de todas las intervenciones aplicadas y materiales utilizados (p.ej. materiales de intervención, materiales de entrenamiento)
- Documentación sobre todos los datos en la computadora y los análisis finales (información sobre las bases de datos, las variables, las salidas de información – ver más adelante).

Obviamente será más fácil juntar todos estos materiales si se prepara cuidadosamente la documentación a medida que trascurre el estudio. Como mínimo cada documento debe llevar la fecha y el nombre de un autor. Los documentos que se guardan en un archivo de procesamiento de texto deben, cuando es posible, contener una anotación de donde se encuentra el archivo del documento, de manera que pueda ser localizado posteriormente para revisión o adaptación para la creación de documentos relacionados.

1.4.9 Curiosea

La cantidad de actividades y de detalles de un proyecto grande puede exceder fácilmente lo que el personal (habitualmente limitado) (e investigadores cortos de tiempo) pueden manejar con comodidad. A pesar de la más alta motivación y experiencia, la comunicación será incompleta y se obviarán temas importantes. Los investigadores deben revisar los formularios de datos regularmente para familiarizarse con los datos en su forma cruda y verificar que la recolección de datos y la codificación se están llevando a cabo en la forma estipulada. Puede valer la pena también hasta curiosear de vez en cuando en los cajones, los montones de formularios, y los archivos de la computadora.

1.4.10 Análisis repetidos de datos

El hecho de que la depuración es una parte tan importante del desarrollo de los programas de computación comerciales sugiere que los investigadores necesitan prever la detección y corrección de errores de programación o por lo menos tratar de minimizar su impacto. Una estrategia es tratar de que diferentes programadores repliquen los análisis antes de su publicación. En general sólo una pequeña proporción de los análisis realizados termina siendo publicada de manera que esta es una estrategia más económica que muchas otras, aunque si hay errores serios que llevaron al análisis en una dirección equivocada, se perderá mucho del trabajo de análisis. La replicación que comienza con los datos lo más crudos posible dan la mayor protección, pero es más frecuente que se usen otros métodos para asegurar que la creación del primer conjunto de datos analizados sea correcto. Se puede obtener una lección objetiva sobre la importancia de la verificación de la precisión de los análisis de datos desde la siguiente cita de una carta al *New England Journal of Medicine* (Ene 14, 1999, p148):

> "En el volumen del 5 de febrero, informamos los resultados de un estudio... Lamentamos comunicar que hemos descubierto un error en la programación de la computadora y que nuestros anteriores resultados son incorrectos... Luego de corregir el error, un nuevo análisis mostró que no hay un aumento significativo..."

Es muy probable que informes de errores como este son sólo la punta del iceberg.

Los errores de especificación o de programación pueden llevar a pérdidas irrecuperables también. En el Ciclo V de la Encuesta Nacional (norteamericana) de Crecimiento de las Familias (en inglés, National Survey of Family Growth, NSFG), por ejemplo, no se obtuvo información sobre si el embarazo era deseado en el momento de la concepción en aproximadamente 5 por ciento de los embarazos, debido a errores en la entrevista personal asistida por computadora (*Public Use Data File Documentation, National Survey of Family Growth, Cycle 5: 1995, User's Guide*, U.S. Department of Health and Human Services, PHS, CDC, National Center for Health Statistics, Hyattsville, Maryland, February, 1997). El primer error ocurrió porque un salto a propósito fue programado (los que respondían que nunca habían tenido relaciones sexuales voluntariamente [variable EVRHDVOL=2] debían saltear la pregunta sobre si el embarazo era deseado) basado en EVRHDVOL≠1. Esto significa que para las mujeres que tenían la variable EVRHDVOL en blanco, también eran salteadas, lo cual era una situación común ya que la pregunta EVRHDVOL no se le hacía a las mujeres que ya habían dicho que su primer relación sexual fue voluntaria.

El segundo error resultó de una combinación de factores que involucraban un salto a propósito de la pregunta sobre el deseo de la concepción cuando la misma resultaba de una relación sexual no voluntaria, basando la implementación de este salto en un intervalo negativo entre las fechas del primer encuentro sexual voluntario y la primer concepción (de manera que parece que la concepción ocurrió previa a la primer relación voluntaria) y una estimación de la fecha de concepción calculada a partir del *mes* en que terminó el embarazo y la duración del embarazo, lo cual producía un resultado que podía ser un mes antes o después de la verdadera fecha. Errores de programación en otros saltos hicieron que 1000 respuestas fueran salteadas en las preguntas sobre abortos espontáneos y que a 2,458 desocupadas equivocadamente no se les hicieran las preguntas sobre su empleo más reciente. Dado que la entrevista NSFG incluía miles de preguntas, la tasa de

error es extremadamente baja. Por supuesto, si los ítems afectados eran los que tú necesitabas analizar estarías desilusionado igual – pero probablemente no tanto como los científicos y administradores cuyas carreras terminaron por la pérdida del primer Aparato de Aterrizaje en Marte debido a un error entre las unidades de medición inglesas y las métricas.

2. Conversión de datos

Imagínate montones de cuestionarios, montones de formularios de resumen de historias clínicas, listas de resultados de laboratorio, fotocopias de los resultados de exámenes, y cosas similares. Antes de que puedan ser analizados, estos datos originales deben ser codificados para ser informatizados (aún si el volumen es suficientemente pequeño y los análisis que pretendemos hacer suficientemente sencillos como para una tabulación manual, igual es necesario codificar). Este proceso puede ser un emprendimiento mayor y arduo, y puede involucrar los siguientes pasos para cada conjunto de valores de datos:

1. Preparación de un manual de codificación que explicite los códigos a utilizarse para cada valor de dato y las decisiones que deben tomarse en todas las situaciones que puedan surgir (ver el manual de codificación de muestra);
2. Codificación de una muestra de formularios de datos para hacer una prueba piloto del manual de codificación (ver cuestionario codificado de muestra);
3. Revisión del manual de codificación, re-codificación de la muestra, y codificación de los restantes formularios (ver las guías de muestra);
4. Mantenimiento de un registro de codificación (ver muestra) donde se relate el número de identificación y las circunstancias para cualquier dato sobre el cual haya surgido una duda o sobre el cual se tomó una decisión fuera de lo habitual, para permitir la revisión y re-codificación si está indicado más adelante;
5. Re-codificación de un porcentaje de los formularios de datos (p.ej., 10%) por parte del personal de supervisión como control de calidad

Dependiendo de la fuente de datos, la codificación puede ser muy rutinaria (p.ej., verificando que una categoría de respuesta ha sido encerrada en un círculo y tal vez anotando el código apropiado para ingresar) o muy exigente (p.ej., la evaluación de una historia clínica para determinar el diagnóstico y el motivo de consulta).

La codificación es una buena oportunidad para revisar cada formulario de datos para observar irregularidades fácilmente detectables, incluyendo la información verbal escrita en el cuestionario. La corrección de las irregularidades (múltiples respuestas a una única pregunta, valores inconsistentes, respuestas que faltan, etc.) generalmente requiere acceso a los formularios de datos, de manera que es más fácil resolver el problema en la etapa de codificación que cuando los formularios han sido archivados y está en uso el conjunto de datos informatizados.

2.1 Limpieza/edición de datos – objetivos

Después de la codificación, los formularios de datos son ingresados con algún tipo de verificación para detectar o corregir los errores de digitación (doble digitación, corrección manual, etc.). Estos datos computarizados deben ser ahora "limpiados" y "editados". La limpieza y la edición de datos también pueden ser consideradas como "control de daño". Es en esta etapa en que se realiza el tamizaje inicial de la información recolectada para evaluar su validez y utilidad. En forma ideal, la limpieza de datos es un proceso permanente; se inicia cuando llegan los primeros resultados. La detección temprana de los errores puede ayudar a minimizarlos en el futuro y asistir en su corrección.

2.1.1 Datos Incompletos

Los datos incompletos son valores que faltan en un único ítem de datos, instrumentos completados parcial o incorrectamente. Un instrumento de encuesta incorrectamente completado puede ser el que tiene el patrón de "saltos" incorrectamente cumplido. La identificación de estos temas y su corrección, cuando es posible, son ambos de importancia.

2.1.2 Valores extremos

Los valores extremos para una variable se denominan "outliers"(valores atípicos). Los valores atípicos pueden también ocurrir para un único lugar (en un estudio multicéntrico) o para un entrevistador que es más "extremo", con respecto a la precisión, el tiempo de entrevista, o a las respuestas. Los valores atípicos pueden cumplir con uno o ambos de dos posibles criterios.

2.1.2.1 Estadistico

Hay pruebas estadísticas formales para los valores estadísticos. Estos procesos están diseñados para identificar aquellos valores que pueden influir de forma inadecuada sobre el análisis estadístico. Una inspección visual de los datos provee, de hecho, mucha información sobre las impresiones que surgen de los potenciales valores atípicos

2.1.2.2 Sustancial

Para que un valor atípico sea verdaderamente un valor atípico, no debe tener un sentido sustancial, por ejemplo una hemoglobina de 0.5 (aunque una hemoglobina de 0.5 puede no ser un valor estadísticamente atípico en un pequeño grupo de pacientes con una anemia severa, cuyo rango esperado puede estar entre 3.5 y 8) o una altura de 9 pies y 10 pulgadas con un peso dado de 95 libras en un adulto aparentemente sano.

2.1.3 Resultados esperados

La observación de los datos con una idea de lo que se espera es útil para determinar que tan "buenos" son estos datos.

2.2 Ubicación del proceso de edición de datos

Alguna disminución de tiempo y distancia entre la recolección de datos y el ingreso al sistema de análisis es útil para la corrección de errores. La edición de los datos debe ocurrir durante todos los momentos de la recolección y análisis de datos. Alguna edición puede ocurrir durante o poco después de la recolección de datos; esto a menudo involucra medios manuales. Procedimientos de edición adicionales ocurrirán más adelante durante la codificación y el ingreso formal de datos. Los procedimientos de edición post-ingreso de datos serán la última etapa del proceso de edición.

2.2.1 Momento de la recolección de datos

Ejemplos de esto son la verificación de la identidad de los sujetos, el uso de números de identificación de los participantes con un dígito verificador, señalización clara de las muestras con etiquetas duplicadas (incluyendo las tapas), revisión precoz de los instrumentos completados, y presentación de instrumentos pre-etiquetados.

2.2.2 Momento del ingreso de los datos (digitación)

Muchos programas de ingreso de datos permiten verificaciones de rangos y/o valores válidos a medida que se ingresan los datos y pueden hasta incluir elementos de verificación de lógica "dura" (consistencia). Las encuestas telefónicas modernas a gran escala usan computadoras para rastrear e ingresar datos durante el proceso de la encuesta. Esto asegura que el instrumento de la encuesta se cumpla correctamente, las respuestas están dentro del rango de tolerancia, y hasta pueden dar una oportunidad para verificación de la consistencia de las respuestas.

2.2.3 Post entrada de datos

La mayor parte de los pasos formales asociados con la edición de datos ocurre después de que los datos han sido digitados y verificados. Involucran el examen de los registros individuales y sus agregados.

2.3 Los pasos del Proceso de Edición

2.3.1 Edición manual

Como comentamos anteriormente, las verificaciones manuales se realizan durante la codificación de los formularios de datos. Esta etapa busca primero que los cuestionarios hayan sido correctamente completados (patrones de salto, etc). La corrección de errores puede significar la necesidad de volver a la fuente original de la información. O en el caso de un resumen de historia clínica (u otro documento) se puede requerir una fotocopia a los efectos de la comparación.

2.3.2 Distribuciones de frecuencia

Las órdenes de SAS PROC FREQ (habitualmente con la opción MISSPRINT o MISSING seleccionada) y PROC UNIVARIATE con las opciones FREQ y PLOT seleccionadas son útiles para examinar las distribuciones de frecuencias. Las distribuciones de frecuencias ayudan a

identificar y cuantificar los datos faltantes, patrones inusuales, y potenciales valores extremos. Por ejemplo, generalmente se registra el peso al nacer en gramos con una precisión de 10 gramos, de manera que el dígito final debe ser cero. La presión arterial generalmente se registra en mmHg, de manera que el dígito final debe distribuirse uniformemente entre 0 y 9. "El caso de los ochos faltantes" " (Stellman SD, *Am J Epidemiol* 129:857-860, 1989) presenta el estudio de un caso en que un analista alerta notó que en la distribución de valores no había ningún valor que contuviera un 8 como dígito terminal. Recién después de mucha verificación e investigación se rastreó el problema a un error de programación (una confusión entre el cero y la letra "o").

2.3.3 Verificaciones lógicas

Estas verificaciones son evaluaciones de comparaciones internas dentro de una única observación.

Se realizan comparaciones "duras" cuando hay inconsistencias obvias, por ejemplo, embarazos en hombres, o después de seguir el proceso de saltos correctamente en el instrumento de la encuesta.

Un ejemplo de una comparación "dura" ocurre cuando el sexo es obtenido de dos fuentes diferentes. Los registros con desacuerdo pueden ser identificados y manejados según el protocolo de estudio. Este protocolo puede indicar que en aquellos casos en que hay desacuerdo se considere que falta el dato, o puede haber una fuente que se considere más confiable (p.ej. el cuestionario que llenó el participante versus la historia clínica), o puede establecerse un nuevo contacto con el participante para verificar el dato. Un desacuerdo entre formularios puede revelar que los formularios pertenecen a distintos participantes.

Son posibles las comparaciones "blandas" pero los valores exactos (de corte) serán dependientes del estudio. El estado civil puede ser interrogado si la edad cuando se produjo el matrimonio está por debajo de cierto valor, que puede ser específico por sexo. La edad exacta se selecciona dependiendo de la población investigada. Otra verificación puede incluir el peso al nacimiento según la edad gestacional.

2.3.4 Presentaciones univariadas

Estas estadísticas son útiles para determinar las medidas de tendencia central (las viejas conocidas media, mediana y modo), las medidas de dispersión (desvío estándar, varianza, rango, percentiles, curtosis y simetría). Además, las representaciones gráficas (p.ej. histograma ,box plot, y gráfico de probabilidad normal) son útiles para describir los datos.

2.3.5 Presentaciones bivariadas

Si los mismos datos han sido recolectados en distintos lugares o momentos, debe evaluarse la concordancia entre las mediciones. Además, las relaciones esperadas pueden ser examinadas: peso al nacer y edad gestacional, presión arterial sistólica y presión arterial diastólica, peso y altura. Combinaciones inusuales deben suscitar una investigación de posibilidad de errores de codificación o digitación.

Las diferencias pueden ser examinadas en el caso de variables continuas y se puede desarrollar un protocolo de aceptación.

2.4 Tratamiento de los valores faltantes

La codificación de las respuestas que faltan o son inconsistentes merece ser pensado cuidadosamente. Un ítem puede no tener respuesta por varias razones, y a menudo es útil distinguir entre estas razones. Por ejemplo, un ítem puede no ser relevante para algunos tipos de participantes (p.ej., una pregunta sobre examen de próstata hecho a una mujer, una pregunta sobre el año en que se usaron por última vez drogas inyectables a una participante que no ha usado drogas inyectables), una pregunta de "tamizaje" puede haber hecho que el participante salteara la pregunta, el encuestado puede no recordar la respuesta, el encuestado puede negarse a responder, o tal vez pueda simplemente omitir una respuesta (en un cuestionario auto administrado) sin dar una razón. Una pregunta como "Se ha realizado una prueba PAP en el último año?" incluida en un cuestionario auto administrado puede no ser contestada por cualquiera de las razones antes mencionadas.

Si faltan muchas respuestas y sólo se utiliza un único código para indicar que la respuesta falta, la distribución de frecuencias dejará al analista preguntándose sobre la utilidad de la pregunta, dado que si hay demasiado pocas respuestas para analizar, la pregunta dará una información limitada. Es preferible usar un código de valor faltante diferente para cada situación. Así una distribución de frecuencias puede mostrar el número de respuestas de cada tipo.

2.4.1 Técnicas para codificar los valores que faltan

Una convención ampliamente utilizada para codificar los valores que faltan es usar códigos numéricos fuera del rango de los valores de la variable, como "999", "998", o "888", para representar los valores faltantes, con un número de dígitos igual a la amplitud del campo. Por ejemplo, en el archivo de datos de uso público para la Encuesta Nacional Norteamericana de Crecimiento de las Familias, respuestas de "no sabe" se codifican como 9, 99, 999, o 9999, las negativas se codifican 8, 98, 998, o 9998, y los valores para "no se comprueba" o "no se obtuvo" son 7,97, 997, o 9997, dependiendo del largo de la columna de los datos originales. Hay varias limitaciones a este procedimiento de codificación. En primer lugar, los paquetes estadísticos pueden no reconocer estos dígitos como valores faltantes, de manera que es fácil, (y vergonzoso) que los valores sean incluidos en los análisis. En segundo lugar, puede ser necesario usar diferentes códigos para la misma razón para faltar debido a los distintos largos de los campos. En tercer lugar, los números no nos dan ningún mecanismo mnemotécnico útil.

Un aspecto muy útil en el Sistema de Análisis Estadístico (Statistical Analysis System, SAS) es aquel que permite códigos especiales para los valores que faltan. Un valor que falta puede ser codificado con uno de 27 distintos códigos para valores faltantes, que consisten en un punto, o un punto seguido por una única letra. Aunque es raro usar más de dos o tres para una variable dada, un análisis podría diferenciar entre "no corresponde", "no sabe" y "se niega" codificando las respuestas como .C, .S y .N respectivamente. Una codificación más elaborada para una variable como las molestias menstruales puede diferenciar entre "no corresponde por el género" (.M), "no corresponde por ser histerectomizada" (.H), y "no corresponde por el cese natural de la menstruación" (.C).

Estos valores pueden ser tabulados por separado en las distribuciones de frecuencias, de manera que la extensión y naturaleza de la "falta" del valor de una variable pueda ser rápidamente evaluada, y el analista no pierde de vista los denominadores. SAS habitualmente no incluye los valores faltantes codificados de esta manera en los cálculos, lo cual ahorra esfuerzos de programación y protege contra errores de programación. La orden TABLES en PROC FREQ da una opción (MISSING) para tratar los valores que faltan igual que todos los demás valores (útil para examinar los porcentajes de valores faltantes) y una opción (MISSPRINT) para mostrar los valores que faltan en las tablas pero no incluirlos en los denominadores para los porcentajes (permitiendo el cálculo correcto de las distribuciones porcentuales para el análisis al mismo tiempo que permite la confirmación sencilla de número de datos y las razones de su ausencia.

2.5 Valores extremos (outliers)

La búsqueda de valores extremos (verificación de rangos) es también un paso preliminar fundamental en el tamizaje de los datos. En primer lugar, los valores extremos deben ser verificados con los datos originales en los formularios de datos para verificar la precisión de la trascripción. Si el valor extremo no puede ser identificado como un error, debe tratarse de evitar distorsionar el análisis.

2.5.1 ¿Que hacer con ellos?

Los valores extremos pueden ser reemplazados con un valor faltante, pero en ese caso se pierde el caso en el análisis (y en un procedimiento de modelado matemático, no se usa el caso entero). Es más, si el valor extremo es un valor legítimo, el hecho de simplemente borrarlo es un procedimiento cuestionable.

Se puede repetir el análisis con y sin el valor extremo para evaluar el impacto de un valor extremo sobre el análisis. O se puede repetir el análisis usando procedimientos estadísticos (no paramétricos) que no son afectados por los valores extremos y los resultados comparados con procedimientos paramétricos – o usar simplemente los procedimientos no paramétricos. Estos procedimientos generalmente involucran medianas, más que medias, o enfocan los rangos de valores de una variable más que los valores en sí mismo. Los procedimientos categóricos en que la variable se clasifica primero en categorías no se verán afectadas por los valores extremos.

2.6 Preocupación con la limpieza/edición de datos

Hay un problema con el manejo de los valores faltantes, los valores extremos y otras verificaciones editoriales: se presta más atención a los problemas extremos y menos atención a los errores que no son tan visibles. Por ejemplo, un error de trascripción que resulta en que un peso al nacer de 2,000 gramos se registre como 3,000 gramos puede no ser detectado para nada una vez que se ha completado la entrada de datos. Pero este sesgo de clasificación errónea puede tener un efecto sustancial si ocurre en una proporción suficientemente grande de las observaciones. Todo el sistema de gestión de datos debe estar diseñado para minimizar y disminuir estos errores. En relación con esta preocupación está la comparación con los valores "esperados". Mientras que esta es una herramienta útil para inspeccionar y comprender los datos, hay una preocupación por tratar de forzar los datos en una distribución esperada. Así se pone el énfasis en los errores de los extremos.

Un error en la dirección opuesta, de más extremo a menos, no es observada con esta definición de verificación de datos. Esta última preocupación se aplica de igual manera en el resto de la verificación de los datos y el análisis.

2.7 Documentacion

La documentación cubre todos los aspectos de los datos además de los problemas identificados, sus soluciones, y todos los cambios hechos a los datos. Algunas técnicas son:

- Mantener una copia maestra del cuestionario y registrar los cambios y las decisions tomadas para cada ítem. Hacer un índice cruzado del cuestionario con el nombre de las variables en los archivos computarizados.
- Mantener por lo menos los datos originales así como todos los programas que llevaron a la creación del último conjunto de datos, de manera que cualquier conjunto de datos intermedio pueda ser recreado si es necesario. (Esta es la justificación para no hacer cambios directos en la base de datos).
- Documentar los programas de computación con un identificador único (i.e., el nombre del programa), título del proyecto, breve descripción del programa, datos que ingresan e información que se obtiene (input y output), cualquier dependencia que pueda tener (programas que DEBAN ser corridos antes del actual o bases de datos esenciales (fecha de la solicitud, persona a quien se solicita, fecha de desarrollo y analista, incluyendo las modificaciones.
- Documentar los programas de computación dentro del programa (en frases de comentarios o titulares) archivos y programas, y en forma externa (i.e. cuadernos).
- Mantener un cuaderno de corridas de programas en orden cronológico, mostrando el nombre (único) del programa, fecha en que se corrió, programador, historia (p.ej. la corrida de nuevo de una versión anterior), conjunto de datos usado, y una descripción de una línea. A veces los programas que crean conjuntos de datos están listados en un sector distinto al de los programas que analizan los datos.
- Trata de usar métodos que se auto-documenten. Adopta un sistema convencional de nombrar los conjuntos de datos, las corridas de computadora, y los nombres de las variables. Selecciona nombres para las variables que tengan sentido si es posible, y permite sufijos (p.ej., en SAS usando 7 caracteres para el nombre de la variable queda un 8º. carácter para designar las re-codificaciones de la variable original). Asigna etiquetas a los conjuntos de variables y a las variables (SAS LABEL u orden ATTRIB) que se guarden dentro de los programas. Si se necesitan más de 40 caracteres, agrega un comentario en el programa que crea la variable o conjuntos de datos. Considera la utilización de etiquetas de valores (formatos) para documentar los valores de cada variable.

3. Análisis de datos

Con la disponibilidad de los paquetes estadísticos para PC, es muy fácil realizar muchas pruebas estadísticas que antes requerían la asistencia de una persona con formación y entrenamiento en bioestadística (y que se distraiga menos de la tarea del análisis de datos), pero hay un aumento del peligro del uso incorrecto, inapropiado o no informado de las pruebas estadísticas (W. Paul McKinney, Mark J. Young, Arthur Hartz, Martha Bi-Fong Lee, "The inexact use of Fisher's Exact Test in six major medical journals" *JAMA* 1989; 261:3430-3433).

Las primeras etapas del análisis deberían poner el énfasis en obtener cierto "conocimiento" de los datos, i.e. alguna familiaridad con sus características esenciales. El proceso de examinar los datos para comprenderlos está integrado al proceso de limpieza y análisis. Siempre cuestiona los datos y examínalos con una visión crítica. Los mismos conceptos que se usaron para la limpieza y edición de datos se pueden aplicar a tratar de comprender los datos. En forma específica, estamos hablando de los valores esperados, los valores que faltan y los valores extremos. Ahora los aplicamos en un "sentido multivariado".

Muchos de los métodos de abordar un conjunto de datos son similares a los descritos anteriormente como limpieza de datos, como la observación de:

1. Distribuciones uni-variables (distribuciones de frecuencias [PROC FREQ], medidas de resumen [PROC UNIVARIATE], gráficas [PROC UNIVARIATE, PROC PLOT u otro].
2. Tabulaciones cruzadas (distribuciones de frecuencia en agrupaciones importantes como sexo, raza, exposición, enfermedad usando PROC FREQ)
3. Gráficos de puntos para mostrar los pares de variables continuas
4. Matrices de correlación

Estos análisis deberían incluir la evaluación de la concordancia cuando se espera que ocurra. Es útil, a menudo, preparar tablas resumen de la información básica del análisis anterior, que puede ser usada como referencia en etapas posteriores de análisis y redacción.

3.1 Resumen de datos

El resumen de datos es una actividad esencial que, como la gestión de datos, ocurre virtualmente en cualquier lugar y momento en que hay datos involucrados. En la fase de análisis de datos, el resumen de datos involucra tomar la decisión de si corresponde y cómo deben agruparse las variables continuas en un número limitado de categorías y si se debe y cómo combinar las variables individuales en escalas e índices. También es necesario crear variables que tengan un mayor sentido conceptual que los ítems individuales.

3.2 Representación grafica

Hay disponibles muchos paquetes para hacer gráficas que permiten graficar, ver, y hasta cierto punto, analizar datos. Las representaciones gráficas de los datos son muy útiles para el estudio de los datos. Los estadísticos a menudo están familiarizados con estas técnicas para examinar los datos, describir los datos y evaluar las pruebas estadísticas (p.ej. gráfico de residuales). El impacto visual de una gráfica da información y permite comprender mejor los datos y disminuye el número de sorpresas que pueden ocurrir. Hay pocos principios generales, dado que cada conjunto de datos es distinto y se enfocará de manera individual. Muchos de los paquetes de gráficas estadísticas modernos disponibles para las computadoras personales tienen una variedad de funciones como el ajuste de curvas, por ejemplo, lineales, cuadráticas, otras curvas polinomiales, y las curvas spline.

3.3 Valores esperados

Tal vez el concepto más importante a tener presente es el de tener una idea de lo que se espera. Este concepto ha sido aplicado durante los procesos de limpieza y edición. El comprender lo que se espera es una función tanto del diseño de estudio como de los valores de los parámetros en la población blanco. Por ejemplo, si se ha utilizado una asignación al azar, los grupos resultantes deben ser similares. Si los controles son seleccionados de la población general a través de la llamada telefónica por discado al azar de dígitos, su perfil demográfico debería reflejar la población entera. Cuando examinamos una tabla, primero verificamos las variables, las etiquetas, y los Ns para la tabla total y las subcategorías que no están incluidas para estar seguro de que comprendes el subconjunto de observaciones que están representadas. En segundo lugar examina las distribuciones marginales para asegurarte que se ajustan a lo que tú esperabas. Luego examina la distribución interna, en particular con respecto al grupo de referencia. Finalmente pasa a evaluar la asociación u otra información de la tabla.

3.4 Valores faltantes

El impacto de los datos faltantes se magnifica en los análisis que involucran gran número de variables, dado que muchos procedimientos analíticos requieren que se omita cualquier observación a la cual le falta el valor para hasta una sola variable en el análisis. De esta manera, si hay cuatro variables, y a cada una le falta el dato en 10% de las observaciones, en el peor caso 40% de las observaciones serían excluidas del análisis. Para evaluar la extensión y la naturaleza de los datos que faltan para una variable, en forma ideal debería hacerse un análisis completo del "valor faltante". Esto significa comparar la presencia/ausencia de información para una variable con otros factores claves, p.ej. edad, raza, género, estado de exposición, y/o estado de enfermedad. El objetivo es identificar las correlaciones de la información faltante. Las relaciones pueden indicar, aunque no en forma conclusiva, que existe un sesgo de selección. Este análisis nos puede dar información sobre como imputar valores para los que faltan (p.ej., el valor de colesterol faltante podría ser estimado en función del sexo, la edad, la raza, y el índice de masa corporal). Relaciones fuertes entre una covariable y los valores faltantes de otra indican que los valores imputados deben ser estratificados por los niveles de la primer covariable.

Aunque reciben relativamente poca atención en las presentaciones introductorias al análisis de datos, los valores faltantes son la maldición de los analistas. El examen de los datos para los

valores faltantes (p.ej., vía SAS PROC FREQ o PROC UNIVARIATE) es un primer paso esencial para cualquier análisis formal. Los códigos especiales de valores faltantes (ver anteriormente) facilitan este examen. Los valores que faltan son una molestia seria o hasta un impedimento en el análisis e interpretación de datos. Una de las mejores motivaciones para diseñar los sistemas de recolección de datos que minimizan los valores faltantes es tratar de lidiar con estos durante el análisis.

3.4.1 Efectos de los datos faltantes

Se pueden distinguir dos tipos de datos faltantes: faltan los datos y faltan los casos. En el primero de los casos, hay información sobre el participante del estudio, pero faltan algunas respuestas. En el caso de que faltan los casos, el participante prospectivo se ha negado a participar o se ha perdido. La presente discusión está dirigida a la situación en que faltan los datos.

Los datos que faltan tienen una variedad de efectos. Como mínimo, los datos faltantes disminuyen el tamaño muestral efectivo, de manera que las estimaciones son menos precisas (tienen intervalos de confianza más amplios) y las pruebas estadísticas tienen menos potencia para excluir la hipótesis nula estadística para las asociaciones observadas. Este problema es agravado en los análisis multivariados (p.ej., análisis estratificado o regresión logística), dado que la mayoría de estos procedimientos eliminan toda observación a la cual le falte un valor en cualquiera de las variables del análisis. Así, un modelo logístico con ocho variables puede perder fácilmente el 30% de las observaciones aún si ninguno de las variables individuales tiene más de 10% de valores faltantes.

Tanto en el análisis univariado como en el multivariado, los datos faltantes llevan a lo que podría denominarse el problema del "denominador cambiante". Cada tabla simple o de doble entrada puede tener distinto número de participantes, lo cual es a la vez desconcertante para los lectores y tedioso para explicar una y otra vez. Una forma de superarlo es la de analizar sólo los casos con datos completos (i.e., observaciones sin valores faltantes), pero el precio a pagar en la pérdida del número de observaciones puede ser totalmente inaceptable.

Las situaciones de datos faltantes se caracterizan en términos del grado y los patrones de las "faltas". Si no hay un patrón sistemático en los datos faltantes para un ítem particular, i.e., todos los participantes tienen la misma probabilidad de omitir una respuesta, los valores faltantes faltan completamente al azar (N.T. en ingles missing completely at random, MCAR). Cuando los datos faltan completamente al azar, las estimaciones a partir de los datos que están presentes no serán sesgadas por los datos faltantes, dado que los datos que no faltan son esencialmente una muestra al azar simple del total (potencial) de datos.

Probablemente es mas frecuente que diferentes grupos de participantes tengan distintas tasas de datos faltantes. En este caso, los datos faltan al azar (N.T. en ingles missing at random, MAR). (suponiendo que los datos faltantes ocurren al azar dentro de cada grupo). Si los grupos que difieren en sus tasas de datos faltantes también difieren en su distribución de la característica que se mide, las estimaciones globales de esa característica también estarán sesgadas.

Por ejemplo, si las personas con múltiples compañeros sexuales tienen mayor probabilidad de negarse a contestar una pregunta sobre ese tema, la estimación del número promedio de compañeros o la proporción de los que responden con más de X compañeros estará sesgada hacia abajo. Las estimaciones de asociaciones con otras variables pueden también estar distorsionada. Es más, los intentos de controlar para esa variable como un factor de confusión potencial puede introducir un sesgo (por quitar selectivamente observaciones del análisis) o debido a un control incompleto del fenómeno de confusión.

3.4.2 ¿Qué se puede hacer con los datos que faltan?

Como en tantas otras áreas de la salud pública, prevenir es mejor. En primer lugar, los formularios y procedimientos de recolección de datos deben ser diseñados y ensayados para tratar de minimizar la falta de datos. En segundo lugar, puede ser posible conseguir una respuesta de un participante vacilante o inseguro (pero dicha inducción debe evitar los peligros de obtener una respuesta imprecisa o que contravenga de alguna manera el derecho del participante a negarse a contestar), contactar de nuevo a los participantes si la revisión del cuestionario muestra una falta de respuestas, u obtener los datos de otra fuente (p.ej., la información que falta en una historia clínica podría obtenerse personalmente del médico tratante). En tercer lugar, puede ser posible combinar los datos de distintas fuentes para crear una variable combinada con menos valores faltantes (p.ej., sexo a partir del cuestionario y sexo a partir de un registro administrativo, aunque puede ser un tema de importancia la precisión diferencial de las fuentes).

A pesar de los mejores esfuerzos, sin embargo, los datos faltantes son cosas de la vida, y es muy raro un estudio observacional que los evita completamente. Sin embargo, cuanto menor el porcentaje de datos faltantes, menor es el problema que crearán y menos importará como los tratemos durante el análisis.

3.4.3 No trates de controlar los valores que faltan de una variable de confusión

Hace unos años se sugirió tratar los valores faltantes como una categoría válida de una variable que es controlada como un potencial factor de confusión. Por ejemplo, si se estaba estratificando una asociación por el hábito de fumar, pueden haber tres estratos: fumador, no fumador, estado desconocido. Trabajos recientes sugieren que esta práctica puede en realidad aumentar el fenómeno de confusión y no es recomendada.

3.4.4 Imputación de los datos faltantes

En los últimos años se ha realizado mucho trabajo para el desarrollo de métodos analíticos para manejar los datos faltantes y minimizar sus efectos perjudiciales. Estos métodos buscan imputar valores para los ítems en que faltan las respuestas de maneras que intentan aumentar la eficiencia estadística (evitando la pérdida de observaciones que tienen uno o pocos valores faltantes) y disminuir el sesgo que resulta cuando los datos que faltan lo hacen al azar (MAR) en vez de completamente al azar (MCAR) (i.e., las tasas de falta de datos varían por subgrupo).

Un método simple de imputación, que ya no se considera adecuado, es el de simplemente reemplazar los valores que faltan con la media o la mediana de las respuestas que hay. Esta práctica

permite que las observaciones con valores faltantes puedan ser utilizadas en análisis multivariados, al mismo tiempo que mantienen la media o mediana global de la variable (calculada con las respuestas presentes). Para variables categóricas, sin embargo, la media puede caer entre categorías (p.ej., la media de una variable con valores 0 y 1, puede ser .3), y para todas las variables sustituyendo con un único valor un gran número de respuestas faltantes cambiará la forma de la distribución de respuestas (aumentando su altura en ese valor y disminuyendo su varianza), con los correspondientes efectos sobre las pruebas estadísticas. Es más, si los valores que faltan no lo hacen completamente al azar, la media de los valores observados puede estar sesgada y por lo tanto también lo estará la media de la variable después de la imputación.

3.4.5 Asignación al azar de los casos que faltan

Un enfoque más sofisticado es el de sacar los valores imputados de una distribución, más que usar un único valor. Así, las observaciones sin valores faltantes (los casos con datos completos) pueden ser usados para generar una distribución de frecuencias para la variable. Esta distribución de frecuencias puede entonces ser usada como base para generar al azar un valor para cada observación a la que le falta una respuesta. Por ejemplo, si la educación fue medida en tres categorías – "menos de secundaria" (25% del conjunto de casos completos), "secundaria completa" (40%), o "más de secundaria" (35%) – entonces por cada observación para la cual falta la educación, se saca un número al azar entre 0 y 1 de una distribución uniforme y se reemplaza el valor que falta con "menos de secundaria" si el número al azar es igual o menor que 0.25, "secundaria completa" si el número fue mayor que 0.25 pero igual a o menor que 0.65, o "más de secundaria" si es mayor que 0.65.

Este método evita introducir una categoría de respuesta adicional y mantiene la forma de la distribución. Pero si los datos que faltan no lo hacen totalmente al azar, la distribución igual será sesgada (p.ej., la mayor no-respuesta de los bebedores intensos disminuirá la estimación del consumo de alcohol; la mayor no-respuesta de los hombres puede también disminuir la estimación del consumo de alcohol).

3.4.6 Imputacion condicional

Los métodos de imputación modernos logran imputaciones más precisas aprovechando las relaciones entre las variables. Si, por ejemplo, las participantes mujeres tienen más probabilidad de tener una confidente que los participantes masculinos, la imputación de un valor para "la presencia de una confidente" se puede basar en el sexo del participante. Con este enfoque, la existencia de confidente entre los hombres se imputará sobre la base de la proporción de hombres con un confidente; la existencia de confidente entre las mujeres se imputará sobre la base de la proporción de mujeres con una confidente. De esta manera, el conjunto de datos que incluye los valores imputados tendrá una estimación menos sesgada de los valores poblacionales que lo que tendría el conjunto de casos completos sólo.

Una extensión sencilla de la imputación condicional de una sola variable es la imputación condicional sobre un conjunto de estratos formado por un número de variables simultáneamente. Si el número de estratos es demasiado grande, se puede usar un procedimiento de regresión para

"predecir" el valor de la variable imputada como función de las variables para las cuales existen datos. Los coeficientes del modelo de regresión se estiman de los casos con datos completos.

Entonces los valores imputados se asignan al azar (usando un procedimiento como el descrito anteriormente) usando las distribuciones específicas por estrato o los valores predichos del modelo de regresión. Esta estrategia da imputaciones superiores para los valores faltantes y mantiene las asociaciones entre las variables imputadas y las otras variables en el modelo o en la estratificación. Cuanto más fuertes las asociaciones entre las variables, más precisa será la imputación. Si en realidad dos variables están asociadas una con la otra, la imputación de un valor a una variable independientemente del valor de la otra hará más débil la asociación.

3.4.7 Imputacion conjunta

Otro paso más adelantado es la imputación conjunta de todos los valores faltantes en cada observación. Imagínate algo que categoriza todas las observaciones con datos completos según los valores de las variables consideradas todas juntas y otro que categoriza todas las demás observaciones según la configuración de los valores faltantes. Supongamos que hay tres variables dicotómicas (0-1), A, B, y C y que A es conocida para todos los participantes pero B y/o C pueden faltar. La planilla puede verse así:

Casos con datos completos

						% de la distribución condicionado a			
Estrato #	A	B	C	Conteo	Porcentaje del total	A	A & B	A, C=0	A, C=1
1	0	0	0	400	33	53	67	80	
2	0	0	1	200	17	27	33		75
							100		
3	0	1	0	100	8	13	67	20	
4	0	1	1	50	4	7	33		25
						100	100	100	100
5	1	0	0	240	20	53	62	83	
6	1	0	1	150	13	33	38		88
							100		
7	1	1	0	40	3	9	67	17	
8	1	1	1	20	2	4	33		12
						100	100	100	100
Total				1,200	100				

Configuracion de los valores faltantes

Configuracion	A	B	C	Conteo
a.	0	0	.	12
b.	0	1	.	18
c.	1	0	.	10
d.	1	1	.	30
e.	0	.	0	40
f.	0	.	1	10
g.	1	.	0	15
h.	1	.	1	25
i.	0	.	.	20
j.	1	.	.	10

En este ejemplo, se enumeran los ocho estratos en la clasificación cruzada de los casos de datos completos del 1 al 8, y los porcentajes para cada estrato son calculados de cuatro maneras diferentes: incondicionalmente (i.e., el conteo como porcentaje de todos los casos con datos completos), condicionado solo al valor de A, condicionado al valor de A y B, y condicionado al valor de A y C [este último requiere dos columnas para que sea más claro]. Mientras tanto las 10 posibles configuraciones con datos faltantes están dispuestas en la segunda tabla y etiquetas a – j.

Luego se lleva a cabo la imputación como sigue. La configuración a. con valor faltante tiene A=0 y B=0, de manera que los 12 casos en esta configuración pertenecen al estrato 1 o el estrato 2. Para mantener esta distribución de los casos con datos completos en esos dos estratos (67% en el estrato 1, 33% en el estrato 2 – ver columna encabezada "A y B"), los 12 casos se asignan aleatoriamente al estrato 1 o al estrato 2 con probabilidades de asignación en esa proporción, de manera que se espera que al estrato 1 le toquen 8 casos y que al estrato 2 le toquen 4 casos. Los 18 casos en la configuración b. tienen A=0 y B=1, de manera que pertenecen al estrato 3 o al estrato 4. Estos 18 casos serán asignados al azar entre los dos estratos con probabilidades proporcionales a la distribución de los casos con datos completos en estos dos estratos (que casualmente resulta igual a los estratos con A=0 y B=0). Las configuraciones c. y d. se tratarán de la misma manera. La configuración e. tiene A=0 y C=0, de manera que los 40 casos de esta configuración pertenecen al estrato 1 o 3. Estos 40 casos se asignaran al azar al estrato 1 o 3 proporcionalmente a la distribución en la columna encabezada "A, C=0". El procedimiento de asignación al azar asignará en promedio 32 casos (80%) al estrato 1, y 8 casos (20%) al estrato 3. Las restantes configuraciones serán manejadas de la misma manera. La configuración i. Tiene A=0 pero no tiene restricciones sobre B o C, de manera que los 20 casos en esta configuración se asignarán al azar entre los estratos 1, 2, 3, o 4 según la distribución en la columna encabezada "A" condicionada a A=0.

La imputación conjunta, condicional hace el máximo uso de los datos disponibles sobre las tres variables, ajusta la distribución de cada variable para dar una mejor estimación de lo esperado para la población entera y mantiene muchas de las asociaciones en doble sentido que involucran las variables imputadas. El procedimiento puede ser llevado a cabo usando un procedimiento de modelado en vez de una clasificación cruzada, que permite la inclusión de más variables.

El procedimiento más actual es el ajuste a un modelo usando el algoritmo de EM ("Expectation Maximization"). El procedimiento BMDP AM usa este procedimiento, pero está diseñado para variables continuas con una distribución normal multivariada e imputa cada variable en forma independiente, de manera que las asociaciones en doble sentido se ven debilitadas. Un nuevo programa de Joe Shafer en la Universidad Estatal de Pennsylvania usa el algoritmo EM con variables categóricas e imputa los datos en forma conjunta; sin embargo requiere recursos informáticos muy potentes.

3.4.8 Imputación múltiple

Todos los procedimientos anteriores resultan en un conjunto único de datos con valores imputados en el lugar de los valores faltantes. Sin embargo, dado que los valores imputados son derivados del resto del conjunto de datos los análisis basados en ellos subestimarán la variabilidad de los datos. Como corrección, el proceso de imputación puede ser llevado a cabo en forma repetida, produciendo múltiples conjuntos de datos, cada uno con un conjunto diferente (aleatorio) de valores imputados. La disponibilidad de múltiples imputaciones permite la estimación de la varianza adicional introducida por el procedimiento de imputación, que puede ser usada entonces para corregir las estimaciones de varianza para el conjunto entero de datos.

[Agradezco a los Dres. Michael Berbaum, Universidad de Alabama en Tuscaloosa y Ralph Foster, Instituto Research Triangle (Carolina del Norte, EEUU) por educarme en este tema y revisar esta sección.]

3.5 Valores extremos

Ahora examinaremos los valores extremos con respecto al enfoque multivariado, i.e. ¿hay valores extremos?. Por ejemplo, estratificas la relación exposición – enfermedad por un factor con 4 niveles. Se observan los 4 odds ratios específicos por estrato 2.3, 3.2, 2.7, y 0.98. El cuarto estrato indica una potencial interacción muy fuerte. ¿Y si este estrato tiene sólo 6 observaciones? Aunque la asociación pueda ser estadísticamente significativa, el colapsar los estratos es razonable dado que la tabla más extrema puede ser el resultado de una imprecisión. En forma alternativa, los valores de la tabla más extrema pueden ser clasificados de nuevo.

3.6 Creación de variables de análisis

Las variables definidas para recoger los datos (como respuestas en un cuestionario, códigos en un formulario de información medica, etc.) no siempre cumplen con los objetivos del análisis. Por ejemplo, un cuestionario sobre comportamientos de riesgo puede utilizar ítems separados para preguntar sobre el uso de crack, cocaína inyectada, heroína inyectada, pero una única variable que combine estos comportamientos ("si" si usa cocaína o heroína, "no" si no usa ninguna de las dos) puede ser más útil para el analista. En ese caso se podría crear una variable derivada (el tratamiento de los valores faltantes se convierte en un tema aquí también). De manera similar, una pregunta sobre estado civil y una pregunta sobre si la persona vive con un "compañero romántico" pueden ser combinadas en una variable que indique "vive con esposo o compañero".

3.7 Decidiendo qué valores incluir en el análisis

No siempre está claro qué valores deben ser incluidos en el análisis. Por ejemplo, habitualmente se excluyen los valores que faltan del denominador para el cálculo de los porcentajes, salvo cuando el objetivo es la evaluación de la gravedad de los valores que faltan. A veces, sin embargo, tiene más sentido tratar por lo menos algunas categorías de valores faltantes de la misma manera que los valores que no faltan. Por ejemplo, una serie de ítems sobre cambios específicos en el comportamiento puede ser precedida de una pregunta de tamizaje, como "¿Has hecho cambios en el último año para disminuir tu riesgo de adquirir VIH?"

Si el participante contesta "si", se le preguntará sobre los cambios específicos; si no, se saltean los detalles específicos. En esta situación, un valor faltante debido al salto de las preguntas específicas en realidad significa "no". Esta situación se puede manejar creando una nueva variable para cada uno de los ítems específicos o re-codificando las variables existentes a "no" cuando la pregunta de tamizaje fue contestada negativamente u otras técnicas. Por el contrario "un verdadero faltante" estaría presente si las preguntas individuales no hubieran sido contestadas aunque la pregunta de tamizaje fue contestada afirmativamente. Este "verdadero faltante" probablemente sería excluido del análisis. De igual manera, si este tipo de cambio de comportamiento no fuera relevante para el participante, el ítem "no corresponde" y probablemente se excluiría la observación también ("probablemente", porque el tratamiento apropiado depende del objetivo del análisis y la interpretación que se le quiera dar).

3.8 Evaluación de los supuestos

Durante esta etapa, se evalúan los supuestos que subyacen las técnicas estadísticas. Por ejemplo, una prueba Chi cuadrada tiene tamaños mínimos esperados en las celdas. Una prueba t supone una distribución gaussiana (normal) en la población. Otros supuestos son aquellos que se realizan sobre la realidad. Por ejemplo, ¿qué pasa si una persona responde a la pregunta sobre su raza marcando tres respuestas Negro, Hispano y Blanco? Hay un protocolo de estudio para clasificar dicho individuo; sin embargo este protocolo puede diferir en otros estudios similares o en el Censo Norteamericano, o en los certificados de nacimiento estaduales, etc. Esto puede tener un impacto en la distribución y/o interpretación esperada.

3.9 Estudio de las preguntas de la investigación

Los análisis de datos pueden ser enfocados en una manera exploratoria o en busca de respuestas a preguntas específicas. En forma ideal, esto último debería haber sido especificado en el protocolo de investigación o por lo menos mucho antes de comenzar el proceso de análisis. A menudo se formulan nuevas preguntas (o todas las preguntas) durante el proceso de análisis. En cualquier caso, es deseable que se articulen preguntas específicas como guía de cómo proceder en el análisis de datos.

Además de su relevancia con respecto a las preguntas, los análisis generalmente deben reflejar el diseño de estudio. Por ejemplo, los diseños de corte no dan estimaciones directas de incidencia, los diseños apareados pueden requerir análisis apareados.

Bibliografía

Calvert, William S. and J. Meimei Ma. *Concepts and case studies in data management* . Cary, NC: SAS Institute, c1996.

Davidson, Fred. Principles of statistical data handling. Thousand Oaks, California, SAGE, 1996, 266pp.

Graham JW, Hofer SM, Piccinin AM. Analysis with missing data in drug prevention research. IN: Collins LM, Seitz LA (eds). Advances in data analysis for prevention intervention research. NIDA Research Monograph 142. U.S. D.H.H.S., N.I.H., National Institute on Drug Abuse, 1994, 13-63.

Marinez, YN, McMahan CA, Barnwell GM, and Wigodsky HS. Ensuring data quality in medical research through an integrated data management system. Statistics in Medicine 1984; 3:101-111.

Hybels, C. Data management outline. Presented at the American Geriatrics Society Summer Workshop. 1989.

Hse J. Missing values revisited. Presented at the all-Merck statisticians conference, October 23, 1989.

Hulley, Stephen B. and Steven R. Cummings. Designing clinical research: an epidemiologic approach. Baltimore, Williams & Wilkins, 1988. Chapter 15: Planning for data management and analysis.

Meinert, Curtis L.; Susan Tonascia. Clinical trials: design, conduct, and analysis. New York, Oxford, 1986.

Raymond, Mark R. Missing data in evaluation research. Evaluation & the health professions 1986;9:395-420.

Spilker, Bert; John Schoenfelder. Data collection forms in clinical trials. Raven Press, 1991.

Sitios Web

Reasearch Data Management, Joachim Heberlein, University of Minnesota, August 28, 1999 has links and case studies concerning 1. federal, University, and disciplinary guidelines governing ownership, access, and retention of research data; 2. choices, decisions, and justifications regarding data (a) accuracy and reliability, (b) ownership, (c) access, (d) use, and (e) retention; 3. preparation of guidelines for the management of research data. www.research.umn.edu/ethics/modResearch2.html

Apéndice

```
*************************************************************************;
*   Los siguientes códigos SAS pueden ser adaptados para crear un carácter de
verificación
```

- para los números de identificación que luego pueden ser usados para detectar errores de trascripción cuando los números de identificación son leídos nuevamente. Por ejemplo, los números de identificación numéricos pueden ser generados por cualquier sistema y luego se le agrega como sufijo o prefijo un carácter de verificación.
- Los identificadores pueden estar impresos en etiquetas para el cuestionario, etiquetas para muestras, formularios de codificación u otros instrumentos para la recolección de datos o el seguimiento. Cuando los números de identificación son digitados con los datos asociados, el programa de entrada de datos puede usar una adaptación del código que sigue para verificar la precisión de la trascripción del mismo número identificador.
- El carácter de verificación generado por el siguiente código detectara errores de trascripción que involucran registrar equivocadamente cualquier digito individual de un número de identificación, inversión de dos dígitos cualesquiera, o hasta errores múltiples, con excepciones muy raras. Dado que los errores en los números de identificación son de los más molestos para detectar y corregir, se recomienda el uso de un carácter de verificación.*
- El código sobre el cual se basa esta rutina SAS fue desarrollado por Robert Thorton en el Instituto Research Triangle (RTI, basado a su vez en un artículo por Joseph A. Gallian ("Assigning Driver's License Numbers", Mathematics Magazine, February 1991, 64(1):13-22). El codigo de Thornton es el método para crear y verificar los números de identificación para el estudio RSVPP. Esta versión de SAS fue desarrollada por Vic Schoenbach, 10/18/94, 10/24/94;

```
Aqui hay una muestra de números de identificación y sus correspondientes digitos
de verificación, tomados de una lista de números de identificación generados por
RTI para el Proyecto RAPP:
*
*         5-1120 -> S  (i.e., El número de identificación completo es 5-1120-S)
*         5-1111 -> T
*         5-1101 -> W
*         5-1011 -> A
*         5-1001 -> D
*         5-1002 -> B
*         5-2001 -> V
*         5-3001 -> Q
*
*************************************************************************;
Este programa lee un listado de números de identificación y les asigna caracteres
de verificación. El programa también lee los caracteres de verificación asignados
por RTI de manera que puedan ser presentados al costado de los caracteres de
verificación calculados para facilitar la verificación de lo correcto de los
caracteres de verificación calculados, con los fines de las pruebas;

data;   * Create a SAS dataset with the original and calculated numbers;

  * Do not write the following variables into the dataset:      ;
    drop alphabet char1 lng sum i mod23 complem ;
```

```
* Define three variables:  the ID number, the check digit, & a 1 byte work area;
attrib strng length=$22 label='ID number needing check digit';
attrib ckd length=$1 label='Check digit to be calculated';
length char1 $1;  * For picking out one character at a time from the ID;
length sum 8;  * For calculation purposes;
alphabet= 'ABCDEFGHIJKLMNOPQRSTUVW';

infile cards;  * Input data file will be on "cards" (i.e., right after
                                                        the program);
input strng $  rti_ckd $ ;  * Read in data (consisting of ID's and the
                   RTI check digit, so that it can be printed in the output);

sum=0;  * Temporary variable to compute running sum;
lng=length(strng);  * Get length of ID to be processed;
if lng > 21 then do;  * Check that the ID is not too long;
        file print;
        put // '*** Error: ' strng= ' is too long = (' lng ')' //;
        file log;  return;  end;

do i = 1 to lng;    * Iterate through each digit of ID number ;
  char1 = substr(strng,lng-i+1,1);   * Extract a character from the ID;
                             * (Hyphens will be ignored);
  if char1 ^= '-' then
     if char1 < '0' or char1 > '9' then do;  * Must be a valid digit - if not
                                               then print error message;
       file print;
       put // '*** Error: Non-numeric character in ID: ' strng= char1=  //;
       file log; return;   * Go back for next ID number;
       end;   * End of then do;
     else do;  * (To get here, character must be a digit from 0-9);
       sum = sum + ((i+1) * char1);  * Take the sum of the digits of the ID
                                  number, weighting each digit by its position;
       end;   * End of else do;
end;    * End of do i = 1 to lng;

* Weighted sum has been obtained - now reduce it;
mod23 = mod(sum,23);  * Calculate the remainder after dividing by 23;
complem = 23 - mod23;  * Take the complement from 23;
ckd=substr(alphabet,complem,1);  * The check character is the
                                   corresponding letter of the alphabet;
return;

cards;   * Here come the test ID's -- note that one is invalid;
   5-1120  S
   5-1111  T
   5-11R1  W     (invalid ID number)
   5-1101  W
   5-1011  A
   5-1001  D
   5-1002  B
   5-2001  V
   5-3001  Q
run;               *(end of list of ID numbers);

* Display the results to verify correctness;
  proc print;  var _all_;  run;
```