EVALUACIÓN DE MÉTODOS ESTADÍSTICOS UTILIZADOS EN TRABAJOS DE GRADO Y TESIS DE LOS PROGRAMAS DE LA FACULTAD DE CIENCIAS AGROPECUARIAS, EN UN PERÍODO DE TRES AÑOS

Hernán Echavarría Sánchez[1]; Guillermo Correa Londoño[2]; Juan Fernando Patiño Díez[3]; Juan José Acosta Jaramillo[4] y Jairo Alberto Rueda Restrepo[5]

RESUMEN

Se hizo un censo de los métodos estadísticos usados en los trabajos de grado y tesis realizados en un periodo de tres años (1999-2001) en la Facultad de Ciencias Agropecuarias de la Universidad Nacional, sede Medellín. En casi la mitad de los trabajos evaluados se encontró al menos un error (49,2 %). A nivel de pregrado, en el programa de Ingeniería Forestal se observó el menor porcentaje de trabajos con al menos un error; mientras que el mayor porcentaje fue observado en Ingeniería Agrícola. La mayoría de los errores se originaron en la poca claridad sobre el papel de la estadística como herramienta para la consecución de los objetivos planteados en los trabajos, lo cual se reflejó en que los trabajos incluyeran resultados estadísticos que en nada contribuían al cumplimiento de los objetivos, en que se omitieran resultados relevantes para su satisfacción y/o en que habiéndose generado resultados pertinentes, no se les diera discusión alguna.

PALABRAS CLAVE: Evaluación de métodos estadísticos, trabajos de grado, tesis.

ABSTRACT

EVALUATION OF STATISTICAL METHODS USED IN FINAL PROJECTS AND THESIS OF THE FACULTAD DE CIENCIAS AGROPECUARIAS, OVER A THREE-YEAR PERIOD

A census of the statistical methods used in the final projects and thesis of the Facultad de Ciencias Agropecuarias of the Universidad Nacional de Colombia, Sede Medellín over a three-year period was done (1999-2001). In almost half of the evaluated works (49,2 %), at least one error was found. At the undergraduate level, the smallest proportion of at-least-one-error works was observed in the

[1] Profesor Asistente. Universidad Nacional de Colombia, Sede Medellín. Facultad de Ciencias Agropecuarias. A.A. 1779, Medellín, Colombia. <hechavar@unal.edu.co>
[2] Profesor Asociado. Universidad Nacional de Colombia, Sede Medellín. Facultad de Ciencias Agropecuarias. A.A. 1779, Medellín, Colombia. <gcorrea@unal.edu.co>
[3] Ingeniero Forestal. Universidad Nacional de Colombia, Sede Medellín. Facultad de Ciencias Agropecuarias. A.A. 1779, Medellín, Colombia. <jfpatino@unalmed.edu.co>
[4] Ingeniero Forestal. Smurfit Kappa Cartón de Colombia S.A. Cl 15 18-109 Puerto Isaacs Yumbo, Cali, Colombia. <juan.acosta@smurfitkappa.com.co>
[5] Instructor Asociado. Universidad Nacional de Colombia, Sede Medellín. Facultad de Ciencias Agropecuarias. A.A. 1779. <jarueda@unal.edu.co>

Recibido: Febrero 24 de 2006; aceptado: Agosto 10 de 2006.

Ingeniería Forestal program, whereas the higher proportion was observed in the Ingeniería Agrícola program. A great bulk of errors was originated in the lack of conscience about the role of statistics as a tool for the achievement of objectives. As a consequence, some works included statistical results that had nothing to do with the objectives; they omitted relevant results and/or failed to discuss them when they appeared.

Key Words: Statistical methods evaluation, Final projects, Thesis.

INTRODUCCIÓN

En los trabajos de grado y tesis, requisitos parciales para obtener el título académico en programas de pregrado y Maestría, respectivamente, los estudiantes hacen uso frecuente de la estadística inferencial y, en menor medida, de la estadística descriptiva, para el análisis y presentación de sus resultados. En cualquier caso, los métodos deben ser seleccionados con base en definiciones precisas de los objetivos específicos, la población objetivo y las restricciones existentes en la toma o generación de la muestra.

La Universidad Nacional de Colombia tiene un compromiso con la comunidad en general, en cuanto a la veracidad del conocimiento que se genera a través de los trabajos de grado y tesis de sus estudiantes; de ahí la importancia de detectar los principales errores conceptuales en que se está incurriendo en los análisis estadísticos que estos conllevan, con el fin de aplicar algunas medidas preventivas que eviten la continuación de tal situación y que den validez externa a su producción académica.

Mediante este trabajo se buscó determinar, para el período evaluado, el porcentaje de trabajos en los cuales el (los) método(s) estadístico(s) utilizado(s) fue(ron) inadecuado(s) e identificar los principales errores conceptuales en que incurrieron los estudiantes al hacer uso de los métodos estadísticos, a fin de proponer los cambios necesarios en los cursos de Bioestadística I (3000009), Bioestadística II (3000010) y Métodos Estadísticos para la Investigación (3000021).

MATERIALES Y MÉTODOS

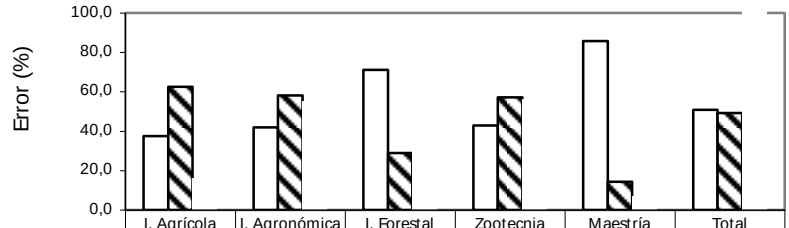
Se realizó un censo de los trabajos de grado y tesis realizados en la Facultad de Ciencias Agropecuarias durante el periodo 1999–2001. No se tomaron en consideración los trabajos realizados en la especialización en Ciencia y Tecnología de Alimentos, pues la exigencia para éstos es de una monografía, la cual generalmente no implica un análisis estadístico. Inicialmente se consideraron 189 trabajos, de los que se analizaron los 179 que usaron métodos estadísticos (172 trabajos de grado y 7 tesis de Maestría), distribuidos así: 16 trabajos de grado de Ingeniería Agrícola, 55 de Ingeniería Agronómica, 45 de Ingeniería Forestal, 56 de Zootecnia y 7 tesis de la Maestría en Bosques y Conservación Ambiental.

Se elaboró una base de datos en la cual se consignó información relacionada con los métodos estadísticos usados en cada uno de los trabajos, se evaluó el uso

adecuado o inadecuado de los mismos, definiendo para cada uno de ellos una lista de errores, carencias y violaciones evidentes a los supuestos; todos éstos fueron denominados con el nombre genérico de "errores". No se incluyó en tal lista el uso inadecuado o impreciso del lenguaje técnico. Seguidamente, se realizó un análisis descriptivo y se discutieron los aspectos más relevantes.

RESULTADOS

Evaluación de métodos estadísticos. En casi la mitad de los trabajos evaluados se detectó al menos un error. En cuanto a pregrados se refiere, en Ingeniería Forestal se observó el menor porcentaje de trabajos de grado con al menos un error; mientras que el mayor porcentaje fue observado en Ingeniería Agrícola. El porcentaje de errores fue similar en los trabajos de grado de Ingeniería Agronómica y en los de Zootecnia; en ambos casos, tal porcentaje es cercano al 60 %. El menor porcentaje global de trabajos con al menos un error se observó en las tesis de la Maestría en Bosques y Conservación Ambiental. En la Figura 1 se discriminan los trabajos por programas con base en la detección de al menos un error.



	I. Agrícola	I. Agronómica	I. Forestal	Zootecnia	Maestría	Total
Sin errores	37,5	41,8	71,1	42,9	85,7	50,8
Con errores	62,5	58,2	28,9	57,1	14,3	49,2
No. trabajos	16	55	45	56	7	179

Figura 1. Porcentaje de trabajos por programa en la Facultad de Ciencias Agropecuarias, con al menos un error, 1999 y 2001.

Identificación de los principales "errores" cometidos. Entre las carencias consideradas genéricamente como "errores", una muy frecuente consistió en indicar el uso de métodos o pruebas, sin incluir ningún tipo de resultado que respaldara la discusión. Asimismo, se consideró inadecuado incluir resultados que no se mencionaron o analizaron en el trabajo y que, en algunos casos, ni siquiera contribuían al logro de los objetivos planteados. Por otro lado, la no especificación del diseño experimental usado o la presencia de

inconsistencias en las tablas de resultados también se clasificó como "error", ya que ello genera dudas acerca del grado de compenetración de los investigadores con la situación estudiada y, por tanto, sobre su capacidad para usar la información disponible de manera adecuada.

Para la mención y posterior discusión de los diferentes errores detectados, éstos se clasificaron de acuerdo al método estadístico en el cual fueron hallados. No se discuten los métodos en los cuales no se detectaron errores, ni aquéllos donde se incluyeron resultados que no se analizaron o donde el error consistió en no presentar resultados del método utilizado.

Medidas de tendencia central. En el 3,25 % de los 154 trabajos en que se obtuvieron, se calculó la media para variables medidas en escala ordinal o, tratándose de variables medidas en escala de razón, se usó la media como medida de tendencia central en una muestra conformada por datos sesgados.

Medidas de dispersión. En 4,84 % de los 124 trabajos en que se obtuvieron, se calculó la varianza para variables medidas en escala ordinal.

Gráficos. En el 1,52% de los 132 trabajos en que utilizó este recurso se usaron histogramas para graficar las frecuencias de variables categóricas.

Intervalos de confianza. En el 19,44 % de los 36 trabajos en los que se usaron, se utilizó la aproximación normal –con un número insuficiente de observaciones– para hallar intervalos de confianza del parámetro *p* de una población binomial. También se calcularon, para variables ordinales, intervalos de confianza para los parámetros μ y σ.

Prueba de hipótesis para dos parámetros. En uno de los seis trabajos que mencionaron este método se utilizó un valor tabular inadecuado para la prueba realizada.

Diseño completamente al azar. En el 66,67 % de los 42 trabajos en que se utilizó este diseño se presentaron errores tales como no aleatorización; no realización de pruebas de comparación de medias a pesar de que el análisis de varianza resultó significativo o viceversa; tras obtener Análisis de Varianza no significativos se procedió a realizar nuevos análisis de varianza sobre sub-conjuntos de tratamientos, con el fin de buscar significancia. En un trabajo se detectó el uso de seudo repeticiones, en otros se usó el diseño completa-mente al azar, a pesar de estar identificada una fuente de variación adicional a los tratamientos. En algunos trabajos se detectaron inconsistencias en las tablas de análisis de varianza presentadas y en un trabajo se hizo una interpretación errónea de los estadísticos calculados.

Diseño de bloques al azar. Entre el 40 % de los 10 trabajos en que se usó este diseño se encontraron los siguientes errores: no aleatorización; se ignoró la estructura factorial de los tratamientos; no se realizaron pruebas de comparación de medias, a pesar de que el Análisis de Varianza resultó significativo, y en un caso se utilizó el diseño experimental inadecuado.

Diseño completamente al azar con arreglo factorial. Entre el 65,22 % de los 23 trabajos en que se usó se encontró como error más frecuente el evaluar efectos principales a pesar de que la interacción entre los factores fue significativa. También se observó que no se analizaron interacciones o se hizo de manera incorrecta; se evaluaron efectos de factores que no fueron significativos en el análisis de varianza; se llego a incluir menos factores de los que realmente se usaron en el experimento; se ignoró la autocorrelación entre los errores de diferentes lecturas tomadas en el tiempo; se evaluaron efectos simples cuando la interacción entre los factores no había sido significativa; se declararon diferencias significativas entre los tratamientos sin haber realizado pruebas de comparación de medias; en un trabajo no se realizaron pruebas de comparación de medias a pesar de que el Análisis de Varianza fue significativo. En otros casos se usó el diseño completamente al azar, a pesar de estar identificada una fuente de variación adicional a los tratamientos; se hizo interpretación errónea de los estadísticos calculados o se detectaron inconsistencias en las tablas de análisis de varianza presentadas.

Diseño en bloques al azar con arreglo factorial. En tres de los seis trabajos que usaron este diseño, el error que más se presentó fue el no analizar las interacciones o hacerlo de manera incorrecta; además se evaluaron efectos principales a pesar de que la interacción entre los factores resultó significativa; se detectó una fuente de variación adicional a los tratamientos y los bloques. En un trabajo no había justificación para el uso de los bloques; no se realizaron pruebas de comparación de medias a pesar de que el análisis de varianza fue significativo y en un trabajo se utilizó como criterio de bloqueo un factor que presentaba interacción evidente con los tratamientos.

Diseño de cuadrados latinos con arreglo factorial. Sólo un trabajo incluyó este diseño pero se ingnoró la estructura factorial de los tratamientos.

Diseño de parcelas divididas en el tiempo. En seis de los 10 trabajos que incluyeron este tipo de análisis se encontraron los siguientes errores: interpretación errónea de los estadísticos calculados; desconocimiento de la autocorrelación entre los errores de diferentes lecturas tomadas en el tiempo; no se evaluó un efecto principal significativo; análisis individuales en el tiempo sin el respaldo de un análisis conjunto en el que se incluyera el tiempo como uno de los factores; conclusiones inconexas con los resultados estadísticos; declaración de diferencias significativas sin el respaldo de las correspondientes pruebas de medias, y en un trabajo el análisis de las interacciones fue incorrecto.

Diseño cross–over. Este diseño se utilizó en cinco trabajos y en tres de ellos se detectaron los siguientes errores: el diseño era inadecuado para la situación y no se realizaron pruebas de comparación de medias, a pesar de haberse obtenido un análisis de varianza significativo.

Prueba de Kruskal–Wallis. En uno de los cuatro trabajos que utilizaron esta prueba, el diseño experimental no era completamente al azar.

Pruebas de comparación de medias. Se realizaron 76 de este tipo de pruebas, de las cuales 31 presentaron algún tipo de error. En este informe se consideraron de manera agrupada pues el error sistemático fue su uso a pesar de que el análisis de varianza no fue significativo. Adicionalmente, en un trabajo se informó haber usado la prueba de Duncan para la evaluación de contrastes y en otro trabajo se detectó una interpretación errónea de las comparaciones realizadas.

Regresión simple. Se detectaron los siguientes errores en 16 de los 36 trabajos en que se uso el método: se usaron modelos de línea recta para modelar asociaciones evidentemente curvilíneas; se detectaron extrapolaciones; se detectaron situaciones en que los residuales presentaban varianzas heterogéneas; en un trabajo se incluyeron resultados que no coincidían con el problema planteado y en otro las conclusiones no correspondían con los resultados estadísticos.

Regresión múltiple. En casi el 46 % de los 24 trabajos que utilizaron este método se encontraron las siguientes fallas: se realizaron extrapolaciones; se presentó un modelo final en el cual algunos de los coeficientes no eran significativos; se detectó una posible autocorrelación de errores pues se hicieron medidas repetidas en el tiempo sobre los mismos individuos y en otro trabajo, se detectó que los residuales presentaban varianzas heterogéneas.

Regresión polinómica. Se detectaron los siguientes errores entre 5 de los 11 trabajos que utilizaron el método: se presentó un modelo polinómico en el cual el término de mayor orden no fue significativo; se hizo extrapolación; en un trabajo las conclusiones no correspondían con los resultados; en otro, se detectó una posible autocorrelación de errores pues se hicieron medidas repetidas en el tiempo sobre los mismos individuos; y en dos de los trabajos, se identificó que los residuales presentaban varianzas heterogéneas.

Correlación. Este tipo de análisis se realizó en 25 trabajos y en el 56 % de ellos se detectaron errores tales como usar los resultados del análisis de correlación como prueba de causalidad y el uso del coeficiente de correlación lineal en una asociación no rectilínea.

Análisis de conglomerados. En cuatro de los 16 trabajos que utilizaron este método se realizaron agrupamientos dentro de un mismo dendrograma con base en diferentes puntos de corte.

Muestreo aleatorio simple. En uno de los ocho trabajos en que se mencionó, se utilizó este esquema muestral, cuando existían estratos claramente identificados.

Muestreo estratificado. En el único trabajo en el que se encontró este esquema muestral se utilizó una fórmula incorrecta para calcular el número de unidades muestrales.

DISCUSIÓN

Una discusión integral de los errores detectados tendría que analizar el porqué de los mismos, lo cual se dificulta por tratarse de eventos relacionados con conductas académicas, más que con fenómenos naturales y por la forma en que se colectó la información. Es posible mencionar algunos eventos que, combinados entre sí o de manera aislada, son determinantes en dicho proceso. El problema puede tener origen en una deficiente formación en la aplicación de los métodos estadísticos, lo cual a su vez podría deberse a un inadecuado enfoque de las correspondientes asignaturas o a la poca conciencia por parte del estudiante de la importancia de tales herramientas para su transcurrir académico en la Universidad y su posterior desempeño profesional. Estos errores podrían, asimismo, estar asociados con la inadecuada o inexistente asesoría estadística, así como con un deficiente seguimiento de las instrucciones suministradas en una adecuada asesoría estadística. A partir de la revisión de los informes finales, no les es posible a los autores de este estudio determinar cuáles de las anteriores circunstancias aplican a cada trabajo.

Además, en los casos en los que las conclusiones no corresponden con los resultados del método estadístico, no es posible discernir si se trata de errores involuntarios por parte del estudiante o de un desconocimiento de la adecuada interpretación de los resultados.

Por lo anterior, la presente discusión se centra en ilustrar las implicaciones de los principales errores detectados y en brindar recomendaciones en muchos de los casos, más que en el porqué de los mismos. En tal sentido, no se discuten los métodos en los cuales no se detectaron errores, ni aquéllos donde el error consistió en incluir resultados que no se analizaron o en no presentar resultados del método utilizado. Esta última situación también fue encontrada por Olsen 2003, en 14 % de los artículos revisados en la revista Infection and Immunity.

El hecho de que en 3 de los 5 programas evaluados, los trabajos con al menos un error superaran los trabajos en los que no se detectaron errores y que en general casi la mitad de los trabajos evaluados (49,2 %) contenía al menos un error, constituye una alerta para estudiantes, directores de trabajos y jurados, pues la veracidad del conocimiento que se generó a través de tales trabajos se vio afectada, no haciéndose un uso óptimo de la información en los casos más favorables, o llegando a conclusiones totalmente incorrectas en los casos más extremos.

Aunque estos resultados son inferiores a los hallados por otros autores en estudios similares, ello no significa que la situación sea menos preocupante, sólo evidencia que no se trata de un problema local. Kanter y Taylor 1994, al revisar 57 artículos de la revista Transfusion, encontraron que el 75 % de los mismos presentaba algún tipo de error estadístico; MacArthur y Jackson 1984, quienes evaluaron 114 artículos del Journal of Infectious Diseases, encontraron que un 95 % de los mismos presentó algún tipo de error estadístico. Los resultados más similares serían los de Olsen 2003, quien

encontró errores en al menos la mitad de los 141 artículos revisados en la revista Infection and Immunity.

Teniendo en cuenta el compromiso de la Universidad con la comunidad y los importantes recursos invertidos en el desarrollo de los trabajos, se hace urgente tomar acciones correctivas que eviten la continuación de tal situación y que den validez externa a la producción académica de la Facultad.

IDENTIFICACIÓN DE LOS PRINCIPALES ERRORES

Estadística descriptiva. Al evaluar el uso de medidas de resumen, se encontró que en algunos trabajos en los que se incluyeron medidas de tendencia central, éstas no se acompañaron con al menos una medida de dispersión. Aunque el uso aislado de una medida de tendencia central no se considera erróneo *per se,* ésta debe ir acompañada por un estadístico de dispersión siempre que ello sea posible, con el fin de describir de manera más completa el conjunto de datos. Esta carencia puede deberse al hecho de que las medidas de tendencia central, a diferencia de las medidas de dispersión, tienen una interpretación más directa, aún por aquellas personas que tomaron sus cursos de estadística hace mucho tiempo.

Los errores relacionados con las medidas de tendencia central y de dispersión se asocian con el uso indiscriminado de la media y la varianza sin tomar en consideración algunas características de los datos, como la escala de medición y la simetría de los mismos.

Al evaluar características medidas en escala ordinal, resulta inadecuado utilizar la media y la desviación estándar como medidas de resumen, pues dado que en este tipo de variables la distancia entre los diferentes niveles o categorías no es constante, la media aritmética no refleja la verdadera tendencia central de los datos. Consecuentemente, la desviación estándar tampoco resulta adecuada, por ser una medida basada en la media. En estos casos, deben usarse la mediana y la desviación mediana como medidas de resumen.

Aunque los estadísticos de forma se utilizan frecuentemente como valores de referencia para inferir sobre la distribución normal de un conjunto de datos, no se recomienda su uso aislado para tal fin; en estos casos es preferible utilizar una prueba formal de bondad de ajuste. Una aplicación adecuada y poco frecuente de los estadísticos de forma, en particular del coeficiente de asimetría, es orientar sobre la medida de tendencia central más adecuada para el conjunto de datos, pues la media, que es la medida de tendencia central más utilizada, sólo resulta adecuada cuando no se presenta una asimetría muy marcada; en caso contrario, se recomienda el uso de la mediana. Por lo anterior, se sugiere usar siempre el coeficiente de asimetría como guía para la elección de la medida de tendencia central (y, en consecuencia, de dispersión) más adecuada.

En lo que a las herramientas gráficas de la estadística descriptiva se refiere, existen dos figuras que por su forma a menudo son confundidas: los histo-

gramas de frecuencias y los diagramas de barras. Los histogramas de frecuencias se usan cuando se tienen variables cuantitativas; en éstos las barras son adyacentes y su ubicación corresponde con el orden intrínseco de la variable. Los diagramas de barras se deben utilizar cuando se tienen variables cualitativas; en éstos las barras están separadas y su ubicación no responde a un orden particular. El uso de un histograma para la caracterización de una variable cualitativa podría generar la falsa idea de continuidad de dicha variable.

Una situación detectada cuando se hacía uso de métodos descriptivos consistió en hacer afirmaciones propias de la estadística inferencial, sin usar las herramientas propias de ésta. Cabe recordar que para generalizar a las poblaciones los comportamientos de las muestras que las representan, no basta con reportar la magnitud de las diferencias observadas; deben usarse métodos inferenciales, los cuales permiten cuantificar la probabilidad de error de las aseveraciones realizadas.

ESTADÍSTICA INFERENCIAL

Estimación. En diversos campos de las Ciencias Agropecuarias, como la producción, el mejoramiento y la modelación de fenómenos naturales, entre otros, con frecuencia se busca estimar parámetros. Si bien, tal estimación puede realizarse de manera puntual, los intervalos de confianza representan un método más completo, pues además de suministrar info.-mación sobre la tendencia central y la dispersión de la variable de interés, se conoce la probabilidad de acierto en la estimación del parámetro. En tal sentido, se considera que los intervalos de confianza deberían usarse más a menudo como complemento a otros métodos; en particular, en la presentación final de los resultados, donde aportarían valiosa información adicional sobre los parámetros de interés, bien sea en el ámbito del diseño de experimentos, la regresión o el muestreo.

El uso de la aproximación normal para obtener intervalos de confianza del parámetro p de una población binomial resulta especialmente inadecuado cuando el tamaño de la muestra es pequeño. En tales casos, lo más adecuado es obtener el intervalo de confianza exacto, usando la distribución binomial. De manera general, dada la actual disponibilidad de herramientas tecnológicas, se recomienda obtener en todos los casos el intervalo de confianza a partir de la distribución exacta.

Tras calcular medidas de resumen para variables ordinales (mediana y desviación mediana), no es adecuado construir intervalos de confianza para los parámetros μ y σ. Lo anterior resulta más evidente si se tiene en cuenta que μ y σ son los parámetros de una distribución normal, la cual resulta inadecuada para modelar fenómenos medidos en escala ordinal. Olsen 2003, también menciona esta situación en artículos médicos, donde se usaron métodos que requerían la distribución normal de los datos y éstos tenían una distrbución sesgada.

Pruebas de hipótesis. El usar valores tabulares inadecuados, ya sea por

errores en el nivel de significancia, en los grados de libertad o en la distribución requerida, hace que las probabilidades reales de error no coincidan con las nominales, lo que conlleva la sobre-estimación o subestimación del riesgo de error o incluso la toma de decisiones incorrectas acerca del juego de hipótesis planteado.

Diseño de experimentos. Se presentan las siguientes consideraciones relaciona-das con los errores detectados en algunos de los trabajos en los que se realizaron experimentos diseñados. Uno de los errores más básicos consiste en usar un diseño experimental inadecuado, bien sea por la no inclusión de fuentes de variación reconocidas o por la inclusión de supuestas fuentes de variación dentro del modelo con el ánimo de disminuir el error experimental o simplemente por tradición.

Un caso particular de la situación planteada anteriormente se presenta cuando el experimento es sometido al esquema de aleatorización restringida propio del Diseño de Bloques al Azar (aleatorización dentro de grupos), sin ningún criterio para ello, manteniendo o retirando los bloques del modelo con base en información a posteriori. En este caso se pueden cometer dos errores: por un lado, evaluar la supuesta significancia de los bloques con base en el valor *p* generado por la mayoría de programas estadísticos, lo cual, según Lenter, Arnold y Hinkelmann, citados por Kuehl 1994 y Petersen 1994, resulta incorrecto, pues no existe una prueba válida para evaluar la significancia de tal efecto. El otro error que puede cometerse en esta situación resulta de retirar el efecto de los bloques del modelo cuando dicho efecto no resulta "significativo", realizando un nuevo análisis de los datos con base en el análisis de varianza de una vía (Diseño Completamente al Azar).

Acorde con Manly 1992, la no aleatorización y/o el uso de seudo repeticiones generan falta de independencia, lo cual afecta la estimación del error experimental y la validez interna del ensayo, propiciando que los efectos de los tratamientos se confundan con efectos de otros factores no considerados en el experimento (factores de confusión). Dos acciones erróneas que hacen que el nivel de significancia real de la prueba supere el nivel de significancia nominal son: realizar pruebas de comparación de medias para un factor cuyo efecto no resultó significativo (Lentner y Bishop 1986, Steel y Torrie 1995) y realizar nuevos análisis de varianza sobre subconjuntos de tratamientos, con el fin de buscar significancia. Estas acciones responden a la creencia común de que en todo ensayo experimental se deben hallar diferencias significativas, desconociendo que la no detección de éstas constituye un resultado igualmente importante para la aproximación a la comprensión del fenómeno estudiado.

El caso contrario ocurre cuando a pesar de que el Análisis de Varianza resulta significativo, no se realizan pruebas de comparación de medias ni contrastes, desconociendo que el análisis de varianza es una prueba general que en ningún caso permite concluir que existen diferencias entre todas las medias; su significancia sólo implica que al menos dos de las medias comparadas

difieren, sin especificar cuáles son las medias causantes de la significancia de la prueba.

En ensayos experimentales que involucran dos o más factores de interés, en ocasiones se ignora la estructura factorial de los tratamientos y se realiza el análisis de varianza que correspondería a una estructura unifactorial. Aunque tal forma de análisis no conlleva errores matemáticos, puede considerarse un error metodológico, pues se desaprovecha la capacidad de tales estructuras de tratamientos para evaluar las posibles relaciones existentes entre los factores evaluados y dificulta, en caso de que tales relaciones existan, presentar recomendaciones específicas para cada uno de los niveles de los factores relacionados. Algo similar ocurre cuando se ignora(n) alguno(s) de los factores involucrados inicialmente en el experimento.

Otro error detectado cuando se realizan experimentos con estructura factorial de tratamientos consiste en evaluar efectos principales, a pesar de que la interacción entre los factores resulte significativa (Petersen 1994). Una interacción implica que las diferencias entre los niveles de un factor cambian, dependiendo de la combinación de niveles de los demás factores. La evaluación de un efecto principal consiste en comparar las medias de cada uno de los niveles de un factor; estas medias se obtienen promediando todas las combinaciones que contengan el nivel de interés, estando allí involucrados todos los niveles de los demás factores. Puesto que, en presencia de interacción, las diferencias entre los niveles de un factor no permanecen constantes a través de cada uno de los niveles de los otros factores, resulta incorrecta la obtención de una diferencia promedio. Por tanto, la evaluación de efectos principales sólo procede cuando la interacción no es significativa y el factor sí lo es. De la misma manera, resulta inadecuado evaluar efectos simples cuando la interacción no es significativa.

En el área biológica son frecuentes los ensayos en los que se realizan lecturas de una variable sobre una misma unidad experimental a través del tiempo. Con ello se genera una estructura de covarianzas entre los errores correspondientes a las diferentes lecturas de una misma unidad experimental. Para el análisis de este tipo de ensayos deben usarse métodos que permitan modelar dicha estructura, la cual no es manejable a través del análisis clásico que asume que las covarianzas entre errores son cero (errores independientes). Para tal efecto, pueden consultarse Littell *et al*. 1998, Littell *et al.* 1996 y SAS 1999, entre otros.

Un procedimiento que ha sido usado por los investigadores para intentar obviar la modelación de la estructura de covarianzas consiste en realizar análisis individuales para cada uno de los tiempos. Debe advertirse que este método no permite evaluar las tendencias de la variable respuesta a través del tiempo, por lo que no constituye un verdadero método de análisis de medidas repetidas y, si bien puede resultar útil para un análisis exploratorio, no es

adecuado como método final de análisis (Littell *et al.* 1998).

Uno de los supuestos del diseño de bloques al azar es la aditividad; esto es, la no existencia de interacción entre bloques y tratamientos. Cuando este supuesto no se cumple, el estimador del error experimental es inadecuado, pues corresponde en realidad al efecto de tal interacción. En estos casos, no es posible estimar el verdadero error experimental, a menos que cada tratamiento esté presente más de una vez dentro de cada bloque (Lentner y Bishop 1986).

REGRESIÓN Y CORRELACIÓN

Regresión simple. Uno de los errores detectados obedece a que existe la creencia de que la significancia de un modelo de regresión lleva implícito el buen ajuste del mismo. Esto, desde luego, no es cierto, ya que la significancia de un modelo de regresión lineal simple sólo indica la presencia de un efecto lineal significativo, pero no da ninguna información acerca de posibles efectos cuadráticos.

Otro error frecuente en el uso de los modelos de regresión es utilizarlo para predecir valores de la variable respuesta a partir de valores de la variable predictora que están por fuera de la región usada para ajustar el modelo. Esta acción no es recomendable, ya que el comporta-miento de un fenómeno en una región determinada no garantiza que dicha tendencia se mantenga por fuera de la misma, lo cual puede causar que las estimaciones así realizadas estén bastante alejadas de la realidad.

Uno de los supuestos para el análisis de regresión es la homocedasticidad. Ésta sólo puede evaluarse de manera objetiva cuando se tienen repeticiones de los valores de la variable independiente. En caso contrario, que es lo más frecuente, este supuesto se evalúa a través del gráfico de residuales estandarizados; aún así, este procedimiento a menudo es omitido. Cuando las varianzas no son homogéneas, no es posible obtener un estimador adecuado de la varianza promedio de las subpoblaciones de la variable dependiente para cada nivel de la variable independiente, razón por la cual la inferencia realizada en tales situaciones resulta inadecuada.

Regresión múltiple. Las consideraciones hechas para los modelos de regresión simple, en general aplican también a los modelos de regresión múltiple. Vale la pena, sin embargo, resaltar algunos aspectos adicionales.

El uso de modelos de regresión múltiple para predecir valores de la variable dependiente a menudo conlleva extra-polaciones, las cuales tienen las mismas implicaciones anotadas para el caso de regresión simple. Su detección, sin embargo, no es tan directa, pues no es suficiente que los valores usados para la predicción estén incluidos en el rango de valores usado para cada una de las variables independientes, sino que ten-dría que verificarse que la combinación evaluada esté en una región que haya sido usada para el ajuste del modelo.

Exceptuando aquellos casos en los cuales deba respetarse el principio de jerarquía, un modelo final de regresión sólo debe contener términos sig-

nificativos, pues de lo contrario se estaría viciando el modelo con relaciones inexistentes, las cuales afectarían la capacidad predictiva del mismo, así como la estimación de los coeficientes de regresión, que miden el grado de asociación entre las demás variables independientes y la variable respuesta.

Para el caso de los modelos de regresión, el principio de jerarquía mencionado anteriormente establece que si un término de orden *k* resulta estadísticamente significativo, el modelo deberá contener todos los términos de orden inferior a *k*, aunque éstos no sean significativos. Este principio también es aplicable a aquellos modelos que incluyen interacciones significativas, en cuyo caso el modelo deberá incluir cada uno de los términos involucrados en la interacción, aun si éstos no son significativos.

Una situación ya mencionada en el contexto de diseño de experimentos, consistente en el uso de medidas repetidas en el tiempo sobre una misma unidad experimental, aparece también en el contexto de regresión, con las mismas implicaciones allí anotadas (en este caso se habla de autocorrelación de errores). Además de verse afectada cada lectura por los valores registrados en los tiempos anteriores, lo que va en contra del supuesto de independencia, los modelos ajustados carecerían de validez externa, es decir, que no podrían utilizarse para describir el comportamiento de la población, pues en realidad se habría modelado el comportamiento de unos individuos.

Correlación. Al evaluar relaciones entre variables a través del coeficiente de correlación, un error consiste en la interpretación de éste como prueba de causalidad, situación errónea pues dicho coeficiente sólo refleja la existencia de asociaciones lineales entre las variables analizadas, sin que de ningún modo la significancia de una correlación implique que el comportamiento de una de las variables sea responsable de los valores que tome la otra.

Como su nombre lo indica, el coeficiente de correlación *lineal* simple sólo es adecuado para caracterizar relaciones lineales entre variables, por lo que resulta erróneo el uso del mismo para caracterizar otros tipos de relaciones. Por lo anterior, al evaluar correlación entre dos variables, se recomienda acompañar el uso del coeficiente de correlación con el correspondiente gráfico de dispersión.

Por lo anterior, la no significancia del coeficiente de correlación -sin el correspondiente diagrama de dispersión- no podrá interpretarse como ausencia de relación entre las dos variables. Sólo podría decirse que no existe relación *lineal* entre las mismas.

ESTADÍSTICA MULTIVARIADA

Al representar una clasificación jerárquica de individuos por medio de dendrogramas, la escogencia de los grupos, para un método de agrupamiento dado, debe estar acorde con la medida de similitud/disimilitud elegida. Si bien pueden plantearse diferentes agrupamientos dependiendo del nivel de detalle establecido, una vez que éste

se elija, deberá usarse para todas las agrupaciones dentro de ese nivel de detalle. No es adecuado, por tanto, al usar métodos de clasificación jerárquica, fusionar subgrupos cuya disimilitud sea mayor que la existente entre otros grupos declarados como diferentes.

Otros. Al plantear el esquema de muestreo para la caracterización de una variable, la presencia de condiciones que puedan afectar su comportamiento, generando subpoblaciones, hace necesario el uso de un esquema de muestreo acorde, esto es, un muestreo estratificado. El uso del esquema de muestreo aleatorio simple en tales casos desaprovecha el conocimiento de tales relaciones, haciendo ineficiente y posiblemente sesgada la caracterización de la variable de interés.

CONCLUSIONES Y RECOMENDACIONES

En este trabajo se definieron una serie de falencias conceptuales, carencias y violaciones evidentes a los supuestos, al aplicar métodos estadísticos. Todos éstos fueron tratados y discutidos con el nombre genérico de "errores". En casi la mitad de los trabajos evaluados (49,2 %) se detectó al menos un error. A nivel de pregrado, en Ingeniería Forestal se observó el menor porcentaje de trabajos con al menos un error; mientras que el mayor porcentaje fue observado en Ingeniería Agrícola. El porcentaje de errores es similar en los trabajos de Ingeniería Agronómica y en los de Zootecnia; en ambos casos, tal porcentaje es cercano al 60 %.

Muchos de los errores se originaron en la poca claridad sobre el papel de la estadística como herramienta para la consecución de los objetivos planteados en los trabajos, lo cual se reflejó en que éstos incluyeran resultados estadísticos que en nada contribuían al cumplimiento de los objetivos, en que se omitieran resultados relevantes para su satisfacción y/o en que habiéndose generado resultados pertinentes, no se les diera discusión alguna.

El principal error detectado en el uso de los métodos de estadística descriptiva consistió en obtener medidas de resumen no adecuadas para la naturaleza de los datos, ya fuera por que éstos presentaban sesgo o porque se midieron en escala ordinal.

En inferencia para una población, se detectó que a pesar de la actual disponibilidad de herramientas tecnológicas que facilitan la obtención del intervalo de confianza exacto para el parámetro p de una población binomial, se sigue obteniendo el intervalo de confianza mediante la aproximación a la distribución normal, lo cual resulta más inadecuado cuando el tamaño de muestra es pequeño.

Los errores encontrados en los trabajos en los que se aplicaron métodos de inferencia estadística para los pará-metros de una y de dos poblaciones, pueden resumirse en uso inadecuado de la aproximación normal. En algunos de los trabajos revisados en esta investigación se obtuvieron intervalos de confianza con base en dicha aproximación, sin tener en cuenta el tamaño de muestra; en otro caso, se usó tal aproximación para

modelar variables ordinales, generando incluso intervalos de confianza para μ y σ. En lo concerniente a los experimentos diseñados los principales errores estuvieron asociados con una aleatorización deficiente o nula, en el uso incompleto de las técnicas del análisis de varianza, dejándose de usar los contrastes de medias para responder las preguntas planteadas en los objetivos, o usando pruebas de medias forzadas (con análisis de varianza no significativo). El mayor porcentaje de errores se observó en experimentos con estructura factorial donde la Interpretación de las interacciones en algunos casos no fue adecuada, pues no se evaluaron efectos simples cuando correspondía o se evaluaron cuando en realidad se debía analizar los efectos principales.

Entre los errores detectados al usar métodos de regresión, los principales estaban relacionados con el no cumplimiento de supuestos tales como homocedasticidad e independencia. Con menor frecuencia se presentaron modelos definitivos con términos no significativos y la extrapolación de conclusiones por fuera del área analizada, además del uso de modelos de línea recta para modelar asociaciones evidentemente curvilíneas.

El principal error detectado en el uso de coeficientes de correlación tuvo que ver con su uso como prueba de causalidad. En un solo trabajo se identificó su uso para caracterizar una relación no rectilínea.

Como recomendación general, se considera necesario concientizar al estudiante sobre la importancia del adecuado dominio de los métodos estadísticos. Este proceso debe emprenderse desde el principio de la carrera, involucrando, no sólo a los miembros de la Sección de Bioestadística, sino a todos los encargados de las asignaturas donde se requiera el uso de tales métodos.

Específicamente, se considera oportuno que en la asignatura Bioestadística I se haga más énfasis en el capítulo de inferencia estadística, especialmente en lo relativo a pruebas de hipótesis, lo cual podría hacerse a expensas de disminuir el énfasis dado a la unidad de álgebra de probabilidades, limitando ésta a la presentación de las definiciones básicas y de las distribuciones de probabilidad más usadas.

En relación a la asignatura Bioestadística II, se considera innecesario continuar incluyendo en el programa el diseño de cuadrado latino, dado su poco uso. En cambio debe incluirse información relacionada con el análisis de medidas repetidas en el tiempo.

Aunque al momento de realizar el censo no existían trabajos de la Maestría en Ciencias Agrarias, se considera que sus tesis podrían exhibir errores similares a los detectados en los demás programas de la Facultad, dado que las interacciones entre estudiantes, profesores, directores y jurados son, asimismo, similares. Por tanto, el curso de Métodos Estadísticos para la Investigación, manteniendo sus particularidades, deberá enfocarse también en prevenir sobre los errores más frecuentemente encontrados.

Se considera muy importante la participación de los directores de trabajos de grado y tesis en una revisión más detallada de la metodología estadística, con la realización de las correspondientes consultas especializadas cuando la situación lo amerite. Con base en los correctivos que se sugieran, podrían optimizarse los recursos invertidos en los trabajos, obteniéndose la mayor información posible que apunte al logro de los objetivos y/o que pueda dar lugar a nuevos proyectos.

RECONOCIMIENTOS

Este trabajo fue realizado con el apoyo financiero de la Dirección de investigación de la Universidad Nacional de Colombia, Sede Medellín –DIME- proyecto 030802611.

BIBLIOGRAFÍA

Kanter, M. H. and Taylor, J. R. 1994. Accuracy of statistical methods in transfusion: a review of articles from July/August 1992 through June 1993. En: Transfusion. Vol. 34, no. 8; p. 697-701.

Kuehl, R. O. 1994. Statistical principles of research design and analysis. Belmont, USA: Duxbury Press. 686 p.

Lentner, M. and Bishop, T. 1986. Expemimental design and analysis. Blacksburg, USA: Valley Book Company. 557 p.

Littell, R. C., Henry, P. R. and Ammerman, C. B. 1998. Statistical analysis of repeated measures using SAS procedures. En: Journal of Animal Science. Vol. 76; p. 1216–1231.

Littell, R. C.; Milliken, G. A.; Stroup, W. W. and Wolfinger, R. D. 1996. SAS® system for mixed models. Cary North Carolina, USA: SAS Institute. 633 p.

Mac Arthur, R. D. and Jackson, G. G. 1984. An evaluation of the use of statistical methodology in the Journal of Infectious Diseases. En: Journal of Infectious Diseases. Vol. 149, no. 3; p. 349-354.

Manly, B. F. J. 1992. The design and analysis of research studies. New York: Cambridge University Press. 353 p.

Olsen, C. H. 2003. Review of the use of statistics in infection and immunity. En: Infection and Immunity. Vol. 71; p. 6689-6692.

Petersen, R. G. 1994. Agricultural field experiments: design and analysis. New York: Marcel Dekker. 409 p.

SAS Institute. 1999. SAS/STAT. Guide for personal computers. Versión 8. Cary, NC: SAS. 378 p.

Steel, R. G. y Torrie, J. H. 1995. Bioestadística: principios y procedimientos. Mexico: McGraw Hill. 622 p.

Zar, J. H. 1984. Biostatistical analysis. New Jersey: Prentice-Hall. 718 p.