Modelamiento del desempeño de los estudiantes universitarios en el primer curso de estadística

Ernesto Peláez García

Universidad del Valle
Facultad de Ingeniería
Escuela de Estadística
Santiago de Cali
2019

Modelamiento del desempeño de los estudiantes universitarios en el primer curso de estadística

Ernesto Peláez García

Trabajo de grado presentado como requisito parcial para optar al título de:

Magíster en Estadística

Director:
Jose Rafael Tovar, PhD.

Facultad de Ingeniería, Escuela de Estadística
Santiago de Cali, Colombia
2019

Dedicatoria

A mi esposa Alexandra, por su amor y paciencia y su constante voz de aliento para que pudiera culminar estos estudios.

A mi madre, la artífice de mi personalidad y valores.

Declaración

Me permito afirmar que he realizado el presente Trabajo de investigación de manera autónoma y con la única ayuda de los medios permitidos y no diferentes a los mencionados en el propio trabajo. Todos los pasajes que se han tomado de manera textual o figurativa de textos publicados y no publicados, los hemos reconocido. Ninguna parte del presente trabajo se ha empleado en ningún otro tipo de Tesis o Trabajo de Grado.

Igualmente declaramos que los datos utilizados en este trabajo están protegidos por las correspondientes cláusulas de confidencialidad.

Santiago de Cali, 19 septiembre de 2019.

_____

Ernesto Peláez García

Agradecimientos

Al profesor Alfonso Bustamante por la consecución del apoyo donde trabajo, para adelantar estos estudios.

A Jennyfer Portilla por su apoyo irrestricto en la modelación estadística de éste trabajo.

A Luis Wallis por su apoyo logístico en la elaboración de éste documento.

Al profesor Rafael Tovar, PhD, por su paciencia con mis tiempos en la dirección de éste trabajo.

Resumen

Los estudiantes universitarios que ingresan a un primer curso de estadística, llegan con preconcepciones, temores, y con algunos conocimientos básicos, si han tenido la oportunidad en los colegios de secundaria, de ver algunos temas relacionados con la estadística. En muchas de las universidades de carácter privado, el fracaso de los estudiantes, originado por afrontar los primeros cursos de matemática universitaria y en este caso en particular, por afrontar un primer curso de estadística, es de gran preocupación. En este sentido la presente investigación quiere indagar, especialmente en alguna de las universidades privadas del sur de Cali, los conocimientos previos, antecedentes académicos, temores, entre otros factores que podrían amenazar el rendimiento y resultado de los estudiantes en un primer curso universitario de estadística, teniendo en cuenta carreras con fuerte componente matemático y las que no lo tienen.

Palabras claves: Desempeño académico, Modelo de regresión ordinal, Estadística Bayesiana, Probabilidad predictiva

Contenido

1 Introducción

El desarrollo de pensamiento estadístico es una de las necesidades más apremiantes en las sociedades modernas, los estudiantes que ingresan a realizar sus estudios de formación superior, en la gran mayoría de los casos, se enfrentan a por lo menos uno o dos cursos de estadística en los que abordan conceptos de probabilidad e inferencia y métodos estadísticos. Se ha observado que los niveles de reprobación y deserción para los cursos de estadística son bastante altos y una buena cantidad de estudiantes aprenden a "odiar" la estadística y lo que se relaciona con la misma.

En este trabajo se estudió la relación entre el desempeño académico (visto como la variable calificación final del curso) con la experiencia previa con conceptos de estadística vistos en la formación básica. El estudio se realizó con 237 estudiantes matriculados en diferentes carreras de una universidad Privada de la Ciudad de Cali. Se aplicaron métodos de regresión para variables de respuesta ordinal y análisis Bayesiano para muestras de la distribución multinomial usando distribuciones previas de Dirichlet.

Entre los resultados más relevantes, se encontró que el tipo de colegio en el que se realizó la formación básica (público o privado) el nivel de matemáticas considerado por el profesor en el curso de estadística y el grado de preocupación del individuo frente a su desempeño en el curso que está por comenzar (medido a través de un índice propuesto por el investigador) son variables que tienen una alta relación en el desempeño final en el curso visto como la nota final obtenida de las evaluaciones realizadas durante el período académico.

2 Planteamiento del problema

Según (Salazar y Jairo, 2017) la importancia de la estadística es tal, que en Colombia, al igual que en muchos otros países, forma parte del currículo de la educación secundaria e incluso se imparten algunos temas en la educación primaria. De acuerdo con (Behar, Ojeda Ramírez, y cols., 2011) la necesidad del mundo actual exige que las personas tengan una formación básica en estadística. Sin embargo, como menciona (Salazar y Jairo, 2017), los estudiantes no reconocen la verdadera utilidad de los conceptos adquiridos y la forma de aplicarlos en su ejercicio profesional, adicional muchos lo asocian con cursos de desarrollo matemático.

Los estudiantes que llegan a estos cursos tienen preconcepciones y actitudes negativas hacia los mismos, que a su vez se asocian con altos niveles de ansiedad al enfrentar el desarrollo del curso, esta percepción puede obedecer a las experiencias personales del estudiante, ya sea el escaso contacto con el área, la forma en que se les enseñó estadística o incluso los comentarios o información tomada de su entorno (Salazar y Jairo, 2017). La actitud que los estudiantes asumen frente a estos cursos puede ser un elemento favorecedor del aprendizaje o, por el contrario, convertirse en una dificultad para finalizar exitosamente el curso. Por tanto, en el proceso de enseñanza - aprendizaje, los profesores encargados de la docencia en esta área ven con preocupación que la predisposición de los alumnos afecte el desarrollo de los cursos (Gil Flores, 1999).

Los profesores sospechan que las formas de enseñanza tradicionales no están dando los resultados deseados y se ha incrementado su preocupación por mejorar la eficacia de sus tareas. Se ha supuesto que la experiencia formativa con asignaturas de matemáticas y de estadística considerando el número de asignaturas cursadas y las notas obtenidas, es el principal determinante de las actitudes de los alumnos, esperando que a mayor experiencia previa o mayor aprovechamiento de la misma habría actitudes más positivas y menor ansiedad por parte de los estudiantes. En este sentido, estudios realizados han mostrado correlaciones positivas entre las actitudes hacia la estadística y variables como el nivel de

conocimientos previos en el área, habilidades matemáticas básicas o número de cursos con contenido estadístico desarrollados previamente (Roberts y Saxe, 1982), han mostrado que existe correlación entre las actitudes y ansiedad de los estudiantes hacia la estadística y el rendimiento en la materia, pero con una capacidad limitada para la predicción de este último (Carmona Márquez, 2004). Al parecer, la formación previa en estadística y matemáticas es un antecedente consistente de las actitudes y la ansiedad hacia la estadística, pero esta relación tiende a ser más débil que la encontrada entre las actitudes o la ansiedad con indicadores de la percepción que tienen los estudiantes acerca de sus capacidades en estas materias, según indica (Carmona Márquez, 2004), por lo que ha propuesto que se tenga en cuenta medidas especificas del aspecto del auto-concepto en investigaciones relacionadas con el tema.

Esta preocupación pone de manifiesto que se abran líneas de investigación sobre los problemas de enseñanza y aprendizaje de la estadística, una de ellas, que permita identificar los factores que inciden en los resultados académicos de los estudiantes en estos cursos. Por esta razón, el siguiente trabajo de investigación pretende analizar el caso de estudiantes que no están matriculados en carreras de estadística o matemáticas pero que deben aprobar al menos un curso como parte de los currículos de sus planes de estudio de pregrado en una universidad privada de la ciudad de Cali. Se espera que los resultados obtenidos a partir del desarrollo de este trabajo de investigación constituyan un aporte significativo a los procesos de evaluación y revisión de currículos de dicha universidad y otros aspectos relevantes.

2.1. Justificación

De acuerdo con (Ponteville, 2014) la estadística posee un papel primordial en el desarrollo de la sociedad actual proporcionando herramientas que permiten describir situaciones de incertidumbre en análisis científicos, sociales y económicos actuales. Analizar la variabilidad, determinar relaciones entre variables, diseñar estudios y experimentos y mejorar las predicciones son algunos de los aspectos que la estadística tiene en cuenta. La adquisición de ideas estadísticas es, por lo tanto, un asunto de gran importancia para la sociedad contemporánea. Como reflejo de este situación, en el caso de las instituciones educativas de nivel superior muchas tienen departamentos académicos de matemática y estadística en forma paralela y la estadística se enseña en departamentos tan variados como los de medicina social, psicología, relaciones del trabajo, entre otros.

La enseñanza de la estadística ha cobrado relevancia desde hace varios años, debido a su importancia en la formación general de las personas (Batanero, Godino, Vallecillos, Green, y Holmes, 1994; Chevallard, 2007), por lo cual, resulta relevante identificar los factores que influyen en el rendimiento académico en un curso introductorio de estadística lo cual permitirá obtener información significativa que contribuya a la construcción de metodologías, a la reestructuración de los programas de los cursos y el planteamiento de metas u objetivos de la formación básica en estadística. De igual forma, los hallazgos se espera contribuyan a reforzar la importancia de una adecuada enseñanza de la estadística en la educación superior, lo que puede ser tenido en cuenta por los organismos administrativos del sistema educativo que trabajan para garantizar una educación con calidad, innovación y pertinencia.

2.2. Objetivos

2.2.1. Objetivo general

Modelar el desempeño de los estudiantes en el primer curso de estadística universitario considerando la información de covariables asociadas a su experiencia previa con los conceptos del área de conocimiento.

2.2.2. Objetivos específicos

- Adaptar un instrumento de medición ya diseñado que permita obtener información acerca del nivel de ansiedad, preconcepciones, y experiencias previas con los conceptos propios de la estadística.
- Desarrollar un índice de preocupación a partir de un conjunto de preguntas del formulario, para incluir la información del mismo en forma de un conjunto de covariables, en un modelo de respuesta ordinal.
- Ajustar un modelo de regresión para variable de respuesta ordinal y evaluar su capacidad clasificadora.

3 Antecedentes

Dentro de los objetivos planteados en este trabajo, está el de la creación de un índice de ansiedad. En la revisión de antecedentes, se expondrá primero una metodología de creación de índices.

(Castaño Vélez y cols., 2010) construyeron un indicador de calidad de vida (ICV) para los hogares del departamento de Antioquia. La metodología estadística empleada en la creación de dicho indicador, se basa en el uso de los métodos de cuantificación de variables cualitativas y el análisis no lineal de componentes principales. (Castaño Vélez y cols., 2010) hace una descripción breve de la metodología estadística que utilizó para la creación del índice mencionado, la cual se expone brevemente a continuación:

"En la construcción de un indicador de calidad de vida, como un resumen de un conjunto de características socio económicas de los hogares, deberían emplearse técnicas estadísticas que permitan transmitir en forma óptima la información del conjunto original de variables al indicador. La optimalidad en este caso consiste en que el indicador debería tener máxima información del conjunto de variables seleccionadas. La selección de estas variables debería ser realizada en términos de su capacidad de discriminar la pobreza."

Si las variables explicativas seleccionadas, son cuantitativas, el ACP (Análisis de Componentes Principales) es el procedimiento estadístico adecuado para construir el índice, el cual se genera como una combinación lineal de las características socio-económicas (o transformaciones de ellas) que es capaz de explicar la mayor parte de la variación total de las variables originales, es decir, que es capaz de conservar la máxima información de dichas variables. Sin embargo, en este caso, muchas de las variables utilizadas en el estudio son de tipo cualitativo, medidas en escala ordinal o nominal, y esta clase de medición no permite la utilización directa del ACP. Una solución a este problema es la transformación de variables cualitativas a variables cuantitativas, lo que significa valorar de alguna manera las categorías de cada una de las variables cualitativas.

A continuación, se presentan algunas investigaciones realizadas sobre factores que influyen en el desempeño académico con el fin de identificar las diferentes técnicas estadísticas empleadas para dar solución al problema planteado y las conclusiones obtenidas.

Ospina y Tapasco (2008) Presentan con base en la implementación del modelo logístico que existe una correlación con el tipo de colegio del cual proviene el estudiante, teniendo menor desempeño en la materia cuando proviene de un colegio oficial vs un colegio privado. En este contexto, se pretende validar que la ansiedad, las expectativas, el tipo de colegio de secundaria de donde egresó el estudiante y otras variables de tipo cualitativo y cuantitativo, afectan su desempeño en el primer curso de estadística universitario y se quiere construir un modelo de predicción que le permita al docente reorientar el curso después de la aplicación del modelo al inicio del mismo.

Garcia (2012) en su trabajo analiza la estructura dimensional de la Escala de Actitudes hacia la Estadística en su aplicación a estudiantes de Ciencias de la Actividad Física y del Deporte. En virtud de los datos obtenidos con una muestra de 145 participantes de ambos sexos que fueron seleccionados por muestreo incidental en dos universidades públicas españolas, se concluye que no son plausibles las estructuras dimensionales propuestas por otros autores. Al mismo tiempo, se defiende una solución factorial basada en tres dimensiones y doce ítems, con capacidad para explicar el 68 % de la varianza del instrumento y con una fiabilidad alfa de Cronbach igual a 0.87. Los estudiantes universitarios de ciencias del deporte declaran niveles medios de ansiedad hacia la estadística, consideran que la utilidad o importancia de esta asignatura es media-baja, y declaran baja predisposición hacia dicha materia.

(Fardales Macias, Diéguez Batista, y Puga García, 2012) analizan las principales fuentes de estrés académico de los estudiantes que ingresan a la universidad y se presenta la validación de un instrumento de evaluación de este constructo psicológico. Los resultados constatan que el estrés académico constituye un fenómeno generalizado en el acceso a la universidad de estudiantes “nuevos” y que los niveles superiores de este estrés se presentan respecto a la exposición de trabajos, la sobrecarga académica, la falta de tiempo y la realización de exámenes. El cuestionario evalúa cuatro dimensiones complementarias (obligaciones académicas, expediente y perspectivas de futuro, dificultades interpersonales y expresión de ideas propias), y muestra una relación significativa de reducida magnitud con el rendimiento

académico y son las mujeres quienes presentan valores superiores en todas las dimensiones evaluadas. Los resultados se discuten en términos de su repercusión para el diseño de sistemas de acción tutorial y de prestación de servicios de atención y asesoramiento psicológico a los estudiantes de nuevo acceso a los estudios universitarios.

(Tovar Cuevas, Castillo, y Marín, 2010) realizaron un estudio de las preconcepciones de estudiantes de la Pontificia Universidad Javeriana sede Cali sobre los cursos de estadística. Éste estudio tuvo como propósito indagar sobre la experiencia de los estudiantes en cursos de estadística y matemáticas, las ideas preconcebidas respecto a los cursos y su conocimiento previo. Se aplicó un cuestionario de 16 puntos a un total de 239 estudiantes. Los resultados mostraron que los estudiantes están de acuerdo en la importancia de realizar al menos un curso de estadística en su entrenamiento personal; y que estos cursos deben pensarse y planearse para que se promueva un ambiente dominado por el pensamiento estadístico, en lugar de ser considerados como cursos de cálculo o de matemáticas en los que los estudiantes hacen cálculos sin tener ideas claras sobre el problema que da lugar a éstos.

(Osorno, 2015) investigó sobre los preconceptos que los estudiantes tienen sobre el primer curso de Estadística en la universidad. El propósito fue explorar aspectos personales y de la formación básica relacionados con la actitud y los imaginarios que tienen los estudiantes de carreras diferentes a Estadística frente al primer curso de estadística. Se aplicó un cuestionario a 816 estudiantes de tres universidades privadas y una pública en Cali. Se propuso un índice de ansiedad en las escalas Importante y Manejable. Se estudió la asociación de los resultados del índice con variables como: Haber cursado estadística en el colegio, facultad de procedencia del estudiante, entre otras. Se identificó que los estudiantes que cursaron estadística en el colegio presentan un índice más alto de ansiedad frente al mismo curso en la universidad. En cuanto, a la facultad se evidenció que algunas facultades como administración, humanidades y salud presentan un índice de ansiedad más alto que en otras.

(Minotta, 2018) propuso un modelo de regresión con variable de respuesta binaria utilizando la función de enlace Log-Log Complementario con el fin de predecir la probabilidad de que un estudiante repruebe o no el curso de estadística (Reprueba con una calificación menor a 3). Según los resultados encontrados, las variables que presentan relación significativa con la probabilidad de reprobar el curso son el desempeño alcanzado en cursos de matemáticas en la universidad, la experiencia en cursos de estadística en el bachillerato, conceptos de

estadística vistos como parte de un curso de matemáticas en el colegio y haber visto al menos cuatro temas de estadística en el bachillerato, con éste modelo se logró un porcentaje correcto de clasificación superior al 50 %. Adicional se realizó una aproximación bayesiana para la probabilidad de reprobar el curso caracterizando a los estudiantes por la combinación de las variables significativas. Se encontró que la mayor probabilidad de riesgo de que los estudiantes reprueben el curso de estadística está en personas que no tuvieron un buen desempeño en cursos de matemáticas en la universidad, tuvieron experiencia en cursos de estadística en el bachillerato, no vieron conceptos de estadística como parte del curso de matemáticas y reportan haber visto cuatro o más temas de estadística en el colegio.

Con base en la revisión de antecedentes, se puede observar que las metodologías estadísticas que podrían ser más consecuentes con el problema estadístico son las de modelamiento para datos de respuesta ordinal y la metodología multivariada de ACP o ACM para la construcción de variables latentes.

4 Marco teórico

Para el desarrollo del trabajo, se realizó la revisión de algunos temas y conceptos necesarios para soportar la metodología desarrollada con miras a obtener los resultados

Se estudiaron algunos conceptos asociados al análisis multivariado de datos. Los cuales fueron usados en la construcción del índice de preocupación frente al curso, y algunos temas relacionados con los modelos de regresión especifica, la regresión con variables de respuesta de tipo ordinal.

4.1. Métodos de análisis multivariado

Los principios de este método se deben a (Guttman, 1941), pero también a (Burt, 1950) y a (Hayashi, 1956). El ACM se utiliza en el análisis de tablas de individuos descritos por variables categóricas y para estudiar las asociaciones entre diferentes modalidades de las variables en estudio. Este análisis parte de una tabla disyuntiva completa $\mathbf{Z}$ de n individuos y p variables categóricas (Pardo y Cabarcas, 2001). En la Figura **4-1** se observa una ilustración de la tabla disyuntiva completa, lo que se observa es que se asigna 1 a la presencia de la característica en la variable y 0 la ausencia de la característica, si se piensa en la relación con el ACS, el ACM es entonces un ACS donde se tiene una variable con n categorías y otra variable con p categorías.

Individuo	Z_A		Z_b			Z_c			$Z_{i.}$
1	1	0	1	0	0	1	0	0	s
2	1	0	1	0	0	1	0	0	s
.	.	.	.	.	.	.	.	.	s
.	.	.	.	.	.	.	.	.	s
n	1	0	0	1	0	0	1	0	s
$Z_{.j}$	$Z_{,1}$	$Z_{,2}$	$Z_{,3}$	$Z_{,4}$	.	.	.	$Z_{.p}$	ns

Tabla **4-1**: Tabla Disjuntiva completa Z_{ij}

- El ACM es considerado una representación gráfica de la asociación entre variables categóricas.

- Los individuos que aparecen cerca se parecen porque asumen las mismas modalidades de diferentes variables.

- Las modalidades de variables diferentes se consideran asociadas porque son asumidas por los mismos individuos.

Ya que es de interés analizar las similitudes entre individuos i e $i^{'}$, se construyen las respectivas distancias ji-cuadrado, las cuales también se pueden construir entre modalidades, y se realizan para la tabla disyuntiva completa Z, dichas distancias están dadas por:

$$z_{im} = \begin{cases} d^2(j,j^{'}) = \Sigma_{i=1}^{n} n \left(\dfrac{z_{ij}}{z_{.j}} - \dfrac{z_{ij^{'}}}{z_{.j^{'}}} \right)^2 \text{ distancia entre las modalidades } j \text{ y } j^{'} \text{ (en } R^n) \\ d^2(i,i^{'}) = \dfrac{i}{k} \Sigma_{j=1}^{p} \dfrac{n}{z_{.j}} \left(z_{ij} - z_{ij^{'}} \right)^2 \text{ distancia entre los individuos } i \text{ y } i^{'} \text{ (en } R^p) \end{cases}$$

En el ACM el objetivo es buscar un nuevo sistema de ejes ortogonales u_α, en los que se proyecte la inercia I_α de la nube de individuos, tal que los primeros ejes concentren la mayor parte de la misma y en forma decreciente.

$$I_\alpha = \psi^{'} M_n^{-1} \psi; \qquad \psi = M_n F M_p u$$

$$I_\alpha = u_\alpha^{'} M_p F^{'} M_n M_n^{-1} M_n F M_p u_\alpha$$

$$I_\alpha = u_\alpha^{'} M_p F^{'} M_n F M_p u_\alpha; \qquad donde \;\; S = F^{'} M_n F M_p$$

$$I_\alpha = u_\alpha^{'} M_p S u_\alpha; \qquad u^{'} M_p u = 1$$

Por tanto en el espacio de los individuos R^p, para encontrar los ejes directores u_α se diagonaliza la matriz $S = F^{'} M_n F M_p$ tal que $u^{'} M_p u = 1$. S No necesariamente es simétrica y por ende no se garantiza que los vectores propios sean ortonormales. Si en vez de diagonalizar S se diagonaliza la matriz S^*:

$$S^* = M_p^{1/2} F^{'} M_n F M_p^{1/2}; \qquad w^{'} w = M_p^{1/2} u$$

$$S^* = S_o^{'} S_o; \qquad S_o = M_n^{1/2} F M_p^{1/2}$$

$$S^* = S_o^{'} S_o; \quad en \quad R_p; \quad w^{'} w = M_p^{1/2} u$$

$$T^* = S_o S_o^{'}; \quad en \quad R_n; \quad r^{'} r = M_n^{1/2} v$$

4.1.1. Inercia Total

La inercia total I no tiene un significado estadístico fácil de interpretar, por que depende del número de modalidades y variables y no de las relaciones entre variables:

$$I = \sum_{q}^{s} I_q = \frac{p}{s} - 1 = \sum_{j=1}^{p} I_j = \sum_{q=1}^{s} I_q$$

4.1.2. Inercia por modalidad

$$I_j = f_{.j} d^2(j, G) = \frac{Z_{.j}}{ns}(\frac{n}{Z_{.j}} - 1) = \frac{1}{s}(1 - \frac{Z_{.j}}{n})$$

La inercia de una modalidad es más grande si la modalidad es rara; es decir si tiene frecuencia muy baja.

4.1.3. Inercia por pregunta

La inercia debida a una pregunta (subtabla) q es función creciente de su número de modalidades p_q

$$I_q = \sum_{j}^{p_q} I_j = \frac{1}{s}(p_q - 1)$$

Se debe equilibrar el número de modalidades por variable y evitar que sean artificialmente activas.

4.1.4. Distribución multinomial

Consideramos un experimento con n ensayos idénticos e independientes, en el que los resultados pueden tomar J valores distintos $(n_1, n_2, ...n_J)$. La probabilidad para cada valor n_j se denota por p_j $j = 1, 2, 3, ..., J$ de modo que $0 < p_j < 1$ y $p_1 + p_2 + ... + p_J = 1$. De esta manera, si N_j es la variable aleatoria que describe el numero de veces que se observa el valor n_j en los n ensayos, para $j = 1, 2, 3, ..., J$, entonces, el vector $N = (N_1, N_2, ..., N_J)$ sigue una distribución multinomial, así:

$$N_1, N_2, ..., N_J | p_1, p_2, ..., p_J \sim Multinomial(n, p_1, p_2, ..., p_J) \tag{4-1}$$

La función de probabilidad de dicha distribución viene dada por:

$$P(N_1 = n_1, N_2 = n_2, ..., N_J = n_J) = \frac{n!}{(n_1!n_2!...n_J!)} p_1^{n_1} p_2^{n_2} ... p_J^{n_J} \tag{4-2}$$

4.1.5. Distribución Dirichlet

La distribución Dirichlet es una extensión multivariante de la distribución beta. Se trata de una distribución ampliamente utilizada en el contexto Bayesiano como una distribución a priori conjugada para una distribución multinomial (Connor y Mosimann, s.f.)

Sea $Y = (y_1, y_2, \ldots, y_j)$ un vector con j componentes, entonces diremos que este vector sigue una distribución Dirichlet con vector de parámetros $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_k)$ es decir,

$$Y|\alpha \sim Dirichlet(\alpha)$$

Y su función de densidad es:

$$P(\alpha_1, \alpha_2, \alpha_3, ..., \alpha_j) = \frac{\Gamma(\sum_{k=1}^{j} \alpha_k)}{\Gamma(\alpha_1)\Gamma(\alpha_2)...\Gamma(\alpha_j)} y_1^{\alpha_1 - 1} y_2^{\alpha_2 - 1} y_3^{\alpha_3 - 1} ... y_j^{\alpha_j - 1} \tag{4-3}$$

En este caso se tiene una distribución conocida para las marginales, que son distribuciones betas con parámetros $a = \alpha_j$ y $b = \alpha_0 - \alpha_j$, para cada una se tiene:

$$E(Y_j) = \frac{\alpha_j}{\alpha_0}, \qquad Var(Y_j) = \frac{\alpha_j(\alpha_0 - \alpha_j)}{\alpha_0^2(\alpha_0 + 1)} \tag{4-4}$$

4.2. Inferencia bayesiana

En la estadística bayesiana los parámetros son considerados variables aleatorias a las cuales se puede asignar una función de distribución que se llamará función de **distribución a priori** y se denotará por $\pi(\theta)$, la cual suele reflejar las creencias sobre los valores posibles que puede asumir el parámetro (Freud, Miller, y M, 2000). Una vez se ha asignado la distribución a priori $\pi(\theta)$ se calcula la distribución posterior de θ que se denota por $\pi(\theta|y_1, y_2, ...)$, de la cual se realizarán todas las inferencias. Usando el Teorema de Bayes, se tiene:

$$\pi(\theta|y_1, y_2, ...) = \frac{L(\theta|y_1, y_2, ...)\pi(\theta)}{\int\limits_{\Theta} L(\theta|y_1, y_2, ...)\pi(\theta)\, d\theta} \tag{4-5}$$

donde Θ es el espacio del parámetro. A menudo la parte más difícil está en calcular la integral en el denominador de la distribución posterior, que se conoce como la constante de normalización. Una forma equivalente de 4-5 es omitir el factor $\int\limits_{\Theta} L(\mathbf{y_1}, \mathbf{y_2}|\theta)\pi(\theta)d\theta$, que no depende de θ sino de los datos de la muestra (Gelman y cols., 1995). De lo anterior se obtiene la distribución a posteri proporcional así:

$$\pi(\theta|y_1, y_2, ...) \propto L(\theta|y_1, y_2, ...)\pi(\theta) \tag{4-6}$$

La distribución posterior $\pi(\theta|y_1, y_2, ...)$ tiene una media, una varianza y cualquier medida, al igual que cualquier otra distribución.

4.2.1. Distribuciones a priori $\pi(\theta)$

Una distribución a priori de un parámetro θ desconocido, es la distribución de probabilidad que expresa el conocimiento acerca de θ antes de tomar en cuenta los "datos". Es posible plantear tantas distribuciones a priori como estados iniciales de conocimiento existan y los diferentes resultados obtenidos en la distribución a posteriori bajo cada uno de los enfoques. Existen diferente tipo de distribuciones a priori:

Distribución a priori no informativa

Distribuciones mínimamente informativas en algún sentido que expresan o representan el estado de ignorancia previa en situaciones en las que no existe información palpable bien sea objetiva o subjetiva, en las que el conocimiento a priori es poco significativo con respecto a

la información muestral. Se dice también que se tiene un conocimiento vago o difuso. Sirven para:

- Deducir creencias a posteriori para quien tiene poco conocimiento previo del asunto.
- Permite la comparación con los resultados de la inferencia clásica que sólo usa la información muestral o parte de ella.
- Estudiar el efecto de la distribución a priori informativa sobre las inferencias al compararlas con las obtenidas usando una a priori de referencia.

Método de Jeffreys

El procedimiento desarrollado por Jeffreys y que se basa en la medida de información de Fisher sobre θ permite tener una distribución a priori con una propiedad de invarianza (Es invariante bajo transformaciones inyectivas). La distribución a priori de Jeffreys es:

Sea $f(x|\theta)$ la densidad de x dado θ. La información de Fisher es definida como:

$$I(\theta) = -E\left(\frac{\partial log(f(x|\theta))}{\partial \theta^2}\right) \tag{4-7}$$

Si θ es un vector con p parámetros entonces:

$$I(\theta) = -E\left(\frac{\partial log(f(x|\theta))}{\partial \theta_i \theta_j}\right) \tag{4-8}$$

Luego la distribución a priori de Jeffreys se define como:

$$\pi(\theta) = \sqrt{I(\theta)} \tag{4-9}$$

En este caso la elección de los valores del vector de hiperparámetros $\alpha = (\alpha_1, \alpha_2, ..., \alpha_k)$ establece el grado de información de la distribución a priori. Las distribuciones a priori que según la revisión de literatura son más utilizadas son:

- Dirichlet$\left(\alpha_1 = \frac{1}{j},\ \alpha_2 = \frac{1}{j},\ ...,\ \alpha_k = \frac{1}{j}\right)$ fue propuesta por (Perks, 1947), es la única distribución a priori que considera el número de categorías, k, como hiperparametro. La cual tiene la siguiente forma:

$$\pi(p_1, p_2, p_3, ..., p_j) = \frac{1}{\prod_{k=1}^{j} \Gamma(\frac{1}{k})} p_1^{\frac{1}{j}-1} p_2^{\frac{1}{j}-1}, p_3^{\frac{1}{j}-1}, ..., p_j^{\frac{1}{j}-1} \tag{4-10}$$

- Dirichlet $\left(\alpha_1 = \frac{1}{2},\ \alpha_2 = \frac{1}{2}\ , ...,\ \alpha_k = \frac{1}{2}\right)$ fue propuesta por Jeffreys, utiliza uno de los procedimientos más comunes para generar distribuciones a priori no informativas y es una de las más globales.

$$\pi(p_1, p_2, p_3, ..., p_j) = \frac{\Gamma(\frac{j}{2})}{\prod_{k=1}^{j} \Gamma(\frac{1}{2})} p_1^{\frac{-1}{2}} p_2^{\frac{-1}{2}}, p_3^{\frac{-1}{2}}, ..., p_j^{\frac{-1}{2}} \tag{4-11}$$

4.2.2. Distribución a posteriori $\pi(\theta|y_1, y_2, ...)$

La distribución a posteriori es la descripción de las creencias del investigador tras la observación de los datos. Una vez conocidos los hiperparámeros de las distribuciones de probabilidad que describen el comportamiento de la cantidad aleatoria de interés (Antes de observar los datos), es decir, en la distribución a priori, se obtiene la forma analítica y gráfica de la distribución de probabilidades actualizada por la información contenida en los datos, a esto se le denomina distribución a posteriori (Tovar Cuevas, 2015).

Distribución predictiva a posteriori

En muchas ocasiones más que describir el comportamiento de un parámetro lo que queremos es describir el comportamiento de observaciones futuras de una variable aleatoria (Predecir) cuya distribución depende de un número finito de parámetros desconocidos, a esta distribución se le conoce como: Distribución predictiva.

Aunque el verdadero valor del parámetro sea desconocido, siempre se dispone de cierta información sobre él mediante la distribución a priori $\pi(\theta)$. Ahora, suponga que se conoce (Hasta cierto punto) el modelo de probabilidad que genera los datos. Asumiendo que los datos $x_1, x_2, \ldots, x_n$ ya fueron observados, que $\pi(\theta)$ es la distribución a priori y que $\pi(\theta|x_1, \ldots, x_n)$ es la distribución a posteriori, entonces la distribución predictiva a posteriori será:

$$f(y_{n+1}|y_1, y_2, ..., y_n) = \int_{\theta} f(y_{n+1}|\theta)\pi(\theta|y_1, ..., y_n)d\theta$$

4.3. Modelo de regresión con respuesta ordinal

Si queremos modelar una variable de respuesta categórica Y, con j categorías con un conjunto de variables explicativas (Factores o Covariables) $X = (X_1, X_2, \ldots, X_m)$ mediante un modelo lineal general, podemos plantearnos lo siguiente:

Sea $\pi_1, \pi_2, .., \pi_j$ representa las probabilidades que aporta cada categoría de la variable respuesta, así:

$$P(y \leq y_j \mid X_1, X_2, ..., X_m)) = \pi_1 + ... + \pi_j; \;\; J = 1, 2, 3, ...j - 1 \tag{4-12}$$

Las probabilidades acumuladas reflejan el orden entre las categorías:

$$P(Y \leq 1) \leq P(Y \leq 2) \leq ... \leq P(Y \leq J)$$

Los logit de las probabilidades acumuladas son:

$$logit(P(y \leq y_j \mid X_1, X_2, ..., X_m)) = log\left(\frac{P(y \leq y_j \mid X_1, X_2, ..., X_m))}{1 - P(y \leq y_j \mid X_1, X_2, ..., X_m))}\right) \;\; J = 1, ..., J - 1 \tag{4-13}$$

El modelo logit de odds proporcionales, cuando se consideran m covariables $X = (X_1, X_2, \ldots, X_m)$ se puede expresar como (Modelo de odds proporcionales):

$$logit(P(y \leq y_j \mid X_1, X_2, ..., X_m)) = \alpha_J - \sum_{i=1}^{m} \beta_i X_i \quad J = 1, 2, 3, .., j - 1 \tag{4-14}$$

$\alpha_J - \sum_{i=1}^{m} \beta_i$ es el predictor lineal y, α_k (Es el parámetro que representa la dependencia con respecto a la categoría j, es también llamado el intercepto) y $\beta = (\beta_1, \beta_2, \ldots, \beta_m)$ son parámetros a estimar. El modelo anterior se basa en el supuesto que los efectos de las variables regresoras son iguales para cualquiera de las categorías y dichos efectos son los mismos para cada α_k.

Las funciones de enlace habituales son logit, clog-log, log-log negativo, probit y Cauchy. Según (Yay y Akıncı, 2009) actualmente no existe un método universal para ayudar al investigador a elegir qué función de enlace se ajusta mejor a un conjunto de datos determinado, sin embargo, en general, el enlace logit se considera adecuado para analizar datos categóricos

ordenados distribuidos uniformemente entre todas las categorías; el enlace clog-log se usa a menudo para analizar datos categóricos ordenados cuando las categorías más altas son más probables mientras que la función de enlace log-log negativa, se usa en casos donde las categorías más bajas son más probables (Cheng, Wang, y Pollastri, 2008). En el caso particular, del presente trabajo se planteará un modelo de regresión ordinal con función de enlace clog-log, ya que, como se verá más adelante la variable respuesta presenta mayor porcentaje para las categorías más altas, el cual se formula, así, los logits acumulativos se forman de la siguiente manera:

$$logit(P(Y \leq y_J \mid X)) = log(-log(1 - P(P(Y \leq y_J \mid X))) = \alpha_J - \sum_{i=1}^{m} \beta_i \quad J = 1, ..., J-1 \tag{4-15}$$

Mientras que una regresión logística binaria modela un logit único, el modelo de regresión logística ordinal modela varios logits acumulativos (Hosmer y Lemeshow, 1989). Por lo tanto, si el resultado ordinal tiene tres niveles (1, 2 y 3), se modelarán dos logits, uno para cada uno de los siguientes puntos de corte: 1 vs 2, 2 vs 3 (Gameroff, 2005).

5 Metodología

5.1. Datos

El conjunto de datos con el que desarrolló la siguiente metodología fue recolectado a través de una encuesta aplicada a estudiantes de la Universidad Icesi que empezarían algún curso de estadística como son: Teoría de probabilidades, Fundamentos de probabilidad y Estadística, Estadística y probabilidad básica, Bioestadística, Bioestadística y fundamentos de epidemiología pertenecientes a seis facultades de la universidad. La encuesta contiene información de distintos aspectos acerca de las expectativas que el estudiante tiene frente al curso de estadística que va a iniciar como también sobre sus conocimientos en la materia y su experiencia frente a este tipo de conocimiento en primaria, bachillerato o universitaria si la tiene. Adicional a los datos recolectados por medio de la encuesta se añadió información sobre la calificación final del curso con el fin de analizar que aspecto puede intervenir en el desempeño final que presenta el estudiante.

5.2. Construcción del índice de Preocupación frente al curso

Para la construcción del índice se propone el uso del método estadístico multivariado análisis de correspondencia múltiple (ACM) ya que las variables que se consideraron para la construcción del mismo son variables cualitativas con categorías. Mediante esta técnica se busca analizar (Si existen) las relaciones entre categorías - individuo que se asocian por tener perfiles parecidos (Lebart, Morineau, y Piron, 1995).

Según (Avella, Oliva, y cols., 2010) el índice generalmente, se construye con la primera componente principal, que se obtiene a través del método multivariado, en este caso particular, el método ACM, el cual puede quedar expresado como:

$\Psi_i = S_D w$
Donde:

Ψ son las coordenadas de los individuos en la primera componente.
$w = M_p^{\frac{1}{2}} u$ tal que $M_p^{\frac{1}{2}}$ es la métrica de las variables y u es el vector propio.
$S_D = M_n^{\frac{1}{2}} F M_p^{\frac{1}{2}}$ tal que $M_n^{\frac{1}{2}}$ es la métrica de los individuos y F es la matriz disjunta estandarizada.

Obtener la primera componente principal por medio de los métodos multivariados, permite representar el índice. Sin embargo, las ponderaciones calculadas de esta manera tienen valores positivos y negativos, lo que hace difícil su interpretación (Avella y cols., 2010). Una forma de reescalar la primera componente principal en una escala de 0 a 100, para representar el índice, se hace por medio de una transformación lineal, así:

$I_i = a\Psi_i + b$
Donde:

$$a = \frac{100}{Max(\Psi_i) - Min(\Psi_i)} \qquad b = -aMin(\Psi_i)$$

Selección de las variables para la construcción del índice

De acuerdo con el objetivo de la creación del índice, se seleccionaron las siguientes variables:

- **¿Cómo espera que sea el grado de dificultad del curso de estadística que matriculó?**

  Fue diseñada para indagar el grado de dificultad que los estudiantes creen encontrar en el curso actualmente matriculado. Esta pregunta tiene cuatro opciones de respuesta: Muy alto, Alto, Bajo y Muy bajo; las dos primeras opciones se podrían relacionar con un nivel de preocupación muy alto o alto, es decir, el nivel de estadística al cual creen que se van a enfrentar es muy alto o alto mientras que las restantes opciones se podrían relacionar con un nivel de preocupación bajo o muy bajo. Para facilitar el análisis esta variable fue categorizada, así: Las opciones Muy alto o alto se agruparon en una categoría llamada **Nivel de dificultad alto** mientras que las opciones Bajo o Muy Bajo fueron agrupadas en la categoría llamada **Nivel de dificultad bajo** (La variable la llamaremos **Dificultad**).

- **¿Cuáles enunciados, podrían ser razón para explicar el grado de dificultad?**

  Fue formulada para indagar las razones por las cuales el estudiante contestó en la pregunta anterior tener un grado de dificultad Muy alto, Alto, Bajo y Muy bajo. Las opciones de respuesta son:

  1. Tengo buenos conocimientos en matemáticas.
  2. Estoy interesado y tengo una buena actitud hacia la carrera.
  3. Tengo confianza porque tuve buenos resultados en los cursos previos.
  4. Me parecen difíciles las materias que utilizan las matemáticas.
  5. No tengo el suficiente tiempo para dedicarle a la materia.
  6. No he visto nada de estadística anteriormente.

  Se debe tener en cuenta que el estudiante puede elegir más de una opción de respuesta, por lo tanto, se decidió categorizar la variable de acuerdo a las siguientes formas de respuesta:

  1. Si seleccionan alguna o todas las opciones 1, 2 y 3 se asume que el estudiante tiene razones positivas para la selección.
  2. Si seleccionan alguna o todas las opciones 4, 5 y 6 se asume que el estudiante tiene razones negativas para la selección.
  3. Si seleccionan opciones de ambos grupos se asume que el estudiante tiene razones positivas y negativas para la selección.

- **¿Cómo se siente ahora frente al curso de estadística que va a comenzar?**

  Fue diseñada para indagar el grado de preocupación frente al curso de estadística actualmente matriculado. Tiene 4 opciones de respuesta: Preocupado, Tenso, Tranquilo y Muy Tranquilo; las dos primeras opciones fueron agrupadas en la categoría "Estoy preocupado" mientras que las restantes categorías fueron Me siento tranquilo.

Construcción de variables para análisis clásico y bayesiano

Inicialmente para llevar a cabo la siguiente propuesta metodológica se debió crear algunas variables con el fin de facilitar el análisis posterior. La primera variable construida fue la calificación final de curso, la cual contó con las siguientes categorías:

$$Calificacion\ final\ del\ curso = \begin{cases} 1 & Menos\ de\ 3 \\ 2\ Entre\ 3\ y\ 4 \\ 3\ Mas\ de\ 4 \end{cases}$$

El estudio cuenta con un total de 237 estudiantes de los cuales 16 quedaron clasificados en la categoría menos de 3, 152 en la categoría entre 3 y 4 mientras que 69 quedaron clasificados en la categoría más de 4.

Posterior a la categorización de la variable calificación final del curso, se procedió categorizar algunas variables que podrían resultar de interés, según los antecedentes consultados y por experiencia de la institución. quedando así:

$$Colegio = \begin{cases} 0 & Privado \\ 1\ Publico \end{cases}$$

La segunda variable considerada fue el índice de preocupación. Este índice fue categorizado en tres grupos, teniendo en cuenta las configuraciones naturales obtenidas a partir de los valores formados por los individuos de la muestra:

$$Indice\ de\ preocupación = \begin{cases} 0 - 20\ Poca \\ 20 - 70\ Media \\ Mas\ de\ 70\ Alta \end{cases}$$

La tercera variable a consideración fue la variable grado del componente de matemáticas en la carrera cursada por el estudiante, esta variable fue construida con información proporcionada por los docentes de la Universidad, considerando las siguientes clasificaciones:

- Los estudiantes que cursaron las asignaturas de Teoría de probabilidades y Bioestadística se consideró que tenían un componente matemático alto en la carrera.

- Los estudiantes que cursaron las asignaturas de Fundamentos de probabilidad y estadística, Estadística y probabilidad básica, Bioestadística y fundamentos de epidemiología se consideró que tenían un componente matemático bajo en la carrera.

$$Grado\ de\ matematicas = \begin{cases} 0 & Alto \\ 1\ Menor \end{cases}$$

Análisis exploratorio

Se realizó la caracterización de las variables: Calificación final del curso, Colegio, Facultad, Índice de preocupación y Grado de matemáticas; por medio de tablas de frecuencias, adicional, se evaluó la asociación entre la variable Calificación final del curso y las restantes variables con el fin de determinar si presentan o no asociación (Se aplicó la prueba de chi-cuadrado para establecer las posibles asociaciones, se usó un nivel de significancia del 0.01 con el fin de incluir en el análisis variables que estén altamente asociadas). Los resultados de estas pruebas permitieron identificar previamente algunos factores que podrían influir el resultado de la variable calificación final del curso.

Modelo de regresión logístico ordinal

Debido a que la variable de interés principal es la calificación obtenida al final del curso, la cual llamaremos: Variable respuesta presenta categorías ordenadas se plantea un modelo de regresión logística ordinal (Dobson, 2012) para identificar las variables regresoras del grupo de Colegio (X_1), Facultad ($X_{21}, X_{22}, X_{23}, X_{24}, X_{25}$), Índice de preocupación (X_{31}, X_{32}) y Grado de matemáticas (X_4) en la carrera matriculada que podría afectar la probabilidad de que un estudiante en particular pertenezca a una de las categorías. Esto es:

$$logit(P(Y \leq y_j \mid X)) = \log(-\log(1 - P(P(Y \leq y_J \mid X)) = \alpha_J - (\beta_1 X_1 + \beta_2 X_{21} + \beta_3 X_{22} + \beta_4 X_{23} + \beta_5 X_{24} + \beta_6 X_{25} + \beta_7 X_{31} + \beta_8 X_{32} + \beta_9 X_4) \quad J = 1, 2$$

Y $P(Y \leq y_j \mid X)$ es la probabilidad acumulada de que un estudiante pertenezca a la categoría j o una menor. Las variables incluidas en el modelo son:

$$X_1 \begin{cases} 0 & Privado \\ 1 & Publico \end{cases}$$

$$X_{21} = \begin{cases} 0 & No\ es\ est.\ Fac.\ de\ Ingenieria \\ 1 & Es\ est.\ Fac.\ de\ Ingenieria \end{cases} \qquad X_{22} = \begin{cases} 0\ No\ es\ est.\ Fac.\ de\ Derecho\ y\ C.S \\ 1\ Es\ est.\ Fac.\ de\ de\ Derecho\ y\ C.S \end{cases}$$

$$X_{23} = \begin{cases} 0\ No\ es\ est.\ Fac.\ de\ Ciencias\ Naturales \\ 1\ Es\ est.Fac.\ de\ \ Ciencias\ Naturales \end{cases}$$

$$X_{24} = \begin{cases} 0 & No\ es\ est.\ Fac.\ de\ Ciencias\ de\ la\ Salud \\ 1 & Es\ est.\ Fac.\ de\ Ciencias\ de\ la\ Salud \end{cases}$$

$$X_{25} = \begin{cases} 0\ \text{No es est. Fac. de Ciencias de la educación} \\ 1\ \text{Es est. Fac. de Ciencias de la educación} \end{cases}$$

$$X_{31} = \begin{cases} 0 & No\ presenta\ Índice\ de\ Preocupación\ media \\ 1 & presenta\ Índice\ de\ Preocupación\ media \end{cases}$$

$$X_{32} = \begin{cases} 0 & No\ presenta\ Índice\ de\ Preocupación\ poca \\ 1 & presenta\ Índice\ de\ Preocupación\ poca \end{cases} \qquad X_4 = \begin{cases} 0\ Alto \\ 1\ Menor \end{cases}$$

En esta formulación, α_j son los parámetros desconocidos (Intercepto) que deben satisfacer la condición de que $\alpha_1 \leq \alpha_2$ y β_i representa el efecto de las variables regresoras sobre las probabilidades acumuladas.

Para obtener una expresión para la probabilidad de que la variable Y (Calificación final del curso) tome una categoría en particular podemos usar la siguiente relación:

$$P(Y \leq 1 \mid X) = P(Y = 1|X) = 1 - e^{-e^{\alpha_j + (\beta_1 X_1 + \beta_2 X_{21} + \beta_3 X_{22} + \beta_4 X_{23} + \beta_5 X_{24} + \beta_6 X_{25} + \beta_7 X_{31} + \beta_8 X_{32} + \beta_9 X_4)))}}$$
$$P(Y = y_2) = P(Y \leq y_2 \mid X) - P(Y = y_1 \mid X)$$
$$P(Y = y_3 \mid X) = 1 - P(Y \leq y_2 \mid X)$$

Para cuantificar la bondad del ajuste global del modelo se calcula la tasa de clasificaciones correctas, es decir, a partir del modelo ajustado, se clasifica cada observación en la categoría más probable, construyendo así una matriz de clasificación observados-predichos y se utiliza el porcentaje de clasificaciones correctas como una medida de la calidad de predicción. Un individuo es clasificado correctamente por el modelo cuando su valor observado de la variable respuesta coincide con su valor estimado por el modelo. Además de construir el modelo y ajustarlo, el siguiente paso será comprobar la significación estadística de cada uno de los coeficientes de regresión en el modelo, a través del estadístico de Wald (Rodriguez,

2008). Adicional se calcula el R^2 de McFadden, el cual nos indicará el grado de ajuste del modelo respecto a si se ajustará un modelo nulo.

Las variables que resulten con un aporte significativo para explicar la probabilidad de que un estudiante en particular pertenezca a una de las categorías, servirán de base para la construcción de los grupos que serán utilizados en el análisis bayesiano.

Modelo bayesiano

De acuerdo con el modelo de regresión logístico ordinal se clasificaron los 237 estudiantes en k grupos dentro de los cuales se asume que en cada uno se ha aplicado el siguiente modelo estadístico. Sea el i-ésimo estudiante que estuvo matriculado en algún curso de estadística durante el primer semestre académico del año 2018, al cual se le ha registrado la calificación final del curso con valores en el intervalo [1 - 5] que llamaremos C de tal forma que se genera la variable Y tal que:

$$Y = \begin{cases} 1 \quad Si\ C < 3 \\ 2\ Si\ C \epsilon [3,4] \\ 3\ Si\ C \geq 4 \end{cases}$$

Se define entonces las variables aleatorias Z_1, Z_2, Z_3 que identifican la categoría de la variable Y obtenida por el estudiante así:

$$Z_j = \begin{cases} 0 \quad Si\ Y \neq j \\ 1 \quad Si\ Y = j \end{cases}$$

con j=1,2,3

Por lo tanto, se puede decir que $Z_j \sim Ber(P_j)$ donde $P_j = P(Z_j = 1)$ con la restricción de que $p_1 + p_2 + p_3 = 1$. Si se tiene una población de **n** estudiantes a los cuales se toma una muestra de n estudiantes entonces:

$n_1 = Identifica\ el\ numero\ de\ los\ n\ estudiantes\ de\ la\ muestra\ que\ obtuvieron\ Z_1 = 1$

$n_2 = Identifica\ el\ numero\ de\ los\ n\ estudiantes\ de\ la\ muestra\ que\ obtuvieron\ Z_2 = 1$

$n_3 = Identifica\ el\ numero\ de\ los\ n\ estudiantes\ de\ la\ muestra\ que\ obtuvieron\ Z_3 = 1$

De tal manera que $n = n_1 + n_2 + n_3$. Es posible asumir que tenemos un vector de observaciones

$n = (n_1, n_2, n_3)$ (Que son realizaciones de las variables aleatorias $n_j = \sum_{i=1}^{n} z_{ij}$) con i= 1,2,...,$n_j$ y un vector de probabilidades $P = (p_1, p_2, p_3)$, y que dadas las características de este experimento (Un estudiante puede quedar clasificado en j categorías, en este caso 3 categorías, mutuamente excluyentes) es posible asumir que el comportamiento natural del vector n puede ser modelado usando una distribución multinomial, cuya función de masa viene dada por:

$$P(n_1, n_2, n_2) = \frac{n!}{(n_1!n_2!n_3!)} p_1^{n_1} p_2^{n_2} p_3^{n_3} \ con\ n_j \geq 0,\ p_1 + p_2 + p_3 = 1,\ n_1 + n_2 + n_3 = n$$

Con función de verosimilitud:

$$L(p_1, p_2, p_3 \mid n_1, n_2, n_3) = \frac{n!}{(n_1!n_2!n_3!)} p_1^{n_1} p_2^{n_2} p_3^{n_3} \ con\ n_j \geq 0,\ p_1 + p_2 + p_3 = 1,\ n_1 + n_2 + n_3 = n$$

El objetivo es realizar la estimación bayesiana de los parámetros p_1, p_2, p_3, enfoque bajo el cual los parámetros son considerados variables aleatorias a las que se puede asignar una función de probabilidad llamada función de distribución a priori denotada por $\pi(p_1, p_2, p_3)$, esta suele reflejar las creencias sobre los valores posibles que puede asumir los parámetros en consideración (Freud y cols., 2000). En este caso se consideró que una *distribución a priori* adecuada para una distribución de muestreo multinomial es la distribución Dirichlet (Distribución conjugada), es decir:

Las distribuciones a priori más utilizadas en la literatura son:

- Distribución propuesta por Perks (1947) en este caso se asume que cada hiperparametro α_j toma el valor de $\frac{1}{j}$, siendo j el número de categorías consideradas en el análisis. La función de la distribución viene dada por:

  $\pi(p_1, p_2, p_3) = \frac{1}{(\prod_{j=1}^{3} \Gamma(\frac{1}{j})} p_1^{\frac{1}{j}-1} p_2^{\frac{1}{j}-1} p_3^{\frac{1}{j}-1}$; $p_1 + p_2 + p_3 = 1$ $, con\ j = 1, 2, ..., c.$ Donde c es el numero de parámetros.

- Distribución propuesta por Jeffrey's (1946) en este caso se asume que cada hiperparametro α_1 toma el valor de $\frac{1}{2}$, esta distribución es obtenida con el método de distribución a priori de Jeffrey's.

  $\pi(p_1, p_2, p_3) = \frac{\Gamma \frac{j}{2}}{(\prod_{j=1}^{3} \Gamma(\frac{1}{j})} p_1^{\frac{1}{j}-1} p_2^{\frac{1}{j}-1} p_3^{\frac{1}{j}-1}$; $p_1 + p_2 + p_3 = 1$

Una vez se ha asignado la distribución a priori $\pi(p_1, p_2, p_3)$ se calcula la distribución posterior de $\pi(p_1, p_2, p_3 \mid n_1, n_2, n_3)$ con la cual se realizarán todas las estimaciones. Usando el teorema de Bayes se tiene:

$$\pi(p_1, p_2, p_3 \mid n_1, n_2, n_3) \propto P(n_1, n_2, n_3 \mid p_1, p_2, p_3)\pi(p_1, p_2, p_3)$$

$$\pi(p_1, p_2, p_3 \mid n_1, n_2, n_3) \propto \frac{n!}{n_1!n_2!n_3!}p_1^{n_1}p_2^{n_2}p_3^{n_3}\frac{\Gamma(\sum_{j=1}^{3}\alpha_j)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)}p_1^{\alpha_1-1}p_2^{\alpha_2-1}p_3^{\alpha_3-1}$$

$$\pi(p_1, p_2, p_3 \mid n_1, n_2, n_3) \propto \overbrace{\frac{n!}{n_1!n_2!n_3!}\frac{\Gamma(\sum_{j=1}^{3}\alpha_j)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)}}^{k}p_1^{n_1}p_2^{n_2}p_3^{n_3}p_1^{\alpha_1-1}p_2^{\alpha_2-1}p_3^{\alpha_3-1}$$

$$\pi(p_1, p_2, p_3 \mid n_1, n_2, n_3) \propto kp_1^{n_1+\alpha_1-1}p_2^{n_2+\alpha_2-1}p_3^{n_3+\alpha_3-1}$$

$$\pi(p_1, p_2, p_3 \mid n_1, n_2, n_3) \propto kp_1^{\alpha_1^*-1}p_2^{\alpha_2^*-1}p_3^{\alpha_3^*-1}$$

Entonces:

$$p_1, p_2, p_3 \mid n_1, n_2, n_3 \sim Dir(\alpha_1^* = \alpha_1 + n_1, \alpha_2^* = \alpha_2 + n_2, \alpha_3^* = \alpha_3 + n_3)$$

Empleando la distribución propuesta por Perks (1947), la distribución a posterior queda dada por:

$$p_1, p_2, p_3 \mid n_1, n_2, n_3 \sim Dir(\alpha_1^* = \tfrac{1}{3} + n_1, \alpha_2^* = \tfrac{1}{3} + n_2, \alpha_3^* = \tfrac{1}{3} + n_3)$$

Empleando la distribución propuesta por Jeffreys (1946), la distribución a posterior queda dada por:
$p_1, p_2, p_3 \mid n_1, n_2, n_3 \sim Dir(\alpha_1^* = \tfrac{1}{2} + n_1, \alpha_2^* = \tfrac{1}{2} + n_2, \alpha_3^* = \tfrac{1}{2} + n_3)$

Luego la distribución predictiva posterior, la cual nos permite conocer la probabilidad de que, al realizar un nuevo estudio dadas las variables seleccionadas y una cantidad igual de estudiantes, que un estudiante sea clasificado en algunas de las categorías es (Calificación menos de 3, entre 3 y 4, más de 4):

Sea $P_l = (p_1, p_2, p_3)$
$P(Y_{n+1} = j \mid n_1, n_2, n_3) = \int_\Delta P(Y_{n+1} = j \mid P_l)\pi(P_l \mid n_1, n_2, n_3)dP_l$
$P(Y_{n+1}j \mid n_1, n_2, n_3) = \int_\Delta P(Y_{n+1} = j \mid P_l) \left[\int P(P_j, P_j \mid n_1, n_2, n_3)dP_{-j}\right] dP_l$
$P(Y_{n+1} = j \mid n_1, n_2, n_3) = \int_\Delta P_{ij}(P_j \mid n_{11}, n_{12}, n_{13})dP_l$
$P(Y_{n+1} = j \mid n_1, n_2, n_3) = \int_\Delta E(P_j \mid n_{11}, n_{12}, n_{13})dP_l$
$P(Y_{n+1} = j \mid n_1, n_2, n_3) = \frac{n_j+\alpha_j}{\sum_{j=1}^{3} n_j+\alpha_j}$

Donde P_{-j} son todos los componentes de P_l excepto P_j.

Empleando la distribución propuesta por Perks (1947), la distribución predictiva a posterior queda dada por:

$$P(Y_{n+1} = j \mid n_1, n_2, n_3) = \frac{n_j+\frac{1}{3}}{\sum_{j=1}^{3} n_j+\frac{1}{3}}$$

Empleando la distribución propuesta por Jeffreys (1946), la distribución predictiva a posterior queda dada por:

$$P(Y_{n+1} = j \mid n_1, n_2, n_3) = \frac{n_j+\frac{1}{2}}{\sum_{j=1}^{3} n_j+\frac{1}{2}}$$

6 Resultados

6.1. Índice de preocupación frente al curso

Después de emplear la metodología en la sección 5.2 se obtuvo una expresión numérica para el grado de preocupación con el que llega el estudiante al primer curso de estadística.

$$I_i = 20{,}5019496(Dificultad_i + ExplicGradoDific_i + SentimCurso_i) + 40{,}26235$$

Donde:

$$Dificultad_i = \begin{cases} 0{,}1944957 & Si\ el\ estudiante\ presento\ dificultad\ Alta \\ -0{,}7274141 & Si\ el\ estudiante\ presento\ dificultad\ Baja \end{cases}$$

$$ExplicGradoDific_i = \begin{cases} 0{,}9002966 & Si\ la\ explicacion\ a\ sus\ razones\ es\ negativa \\ -0{,}5961097 & Si\ la\ explicacion\ a\ sus\ razones\ es\ positiva \\ 1{,}7100800 & Si\ la\ explicacion\ de\ sus\ razones\ son\ positivas\ y\ negativas \end{cases}$$

$$SentimCurso_i = \begin{cases} 1{,}0091789 & Si\ el\ estudiante\ se\ siente\ Ansioso \\ -0{,}6403066 & Si\ el\ estudiante\ se\ siente\ tranquilo \end{cases}$$

La distribución de los valores del índice en los estudiantes del estudio se aparecen en la figura 6-1:

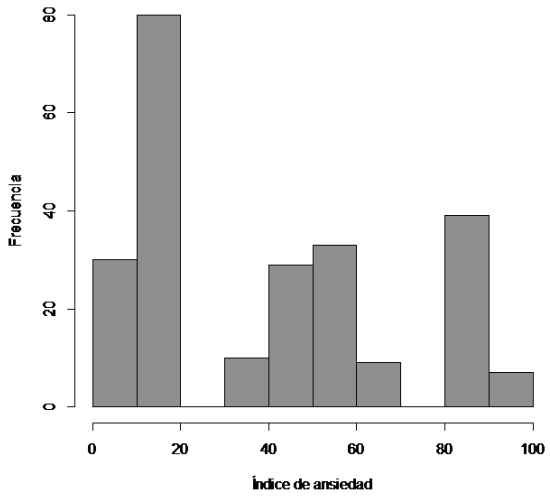

Figura **6-1**: Histograma del índice de preocupación

Con base en la forma de la distribución en el índice se categorizó como:

$$\textit{Índice de preocupación} = \begin{cases} 0-20 \;\; Poca \\ 20-70 \;\; Media \\ Mas\ de\ 70\ Alta \end{cases}$$

En la tabla **6-1** aparece como quedan clasificados los estudiantes de acuerdo con los valores que tomó el índice de preocupación de acuerdo con las categorías que toman las variables **Grado de dificultad de curso, Razones del grado de dificultad y Sentimiento frente al curso.**

Tabla **6-1**: Tabla de clasificación de estudiantes según el índice de preocupación

Grado de dificultad del curso (GradoDificEsp)	**Razones del grado de dificultad (ExplicGradoDific)**	**Sentimiento frente al curso (SentimCurso)**	**Clasificación del Índice de preocupación**
Alta	Negativas	Ansioso	Alta
Alta	Negativas	Tranquilo	Media
Alta	Positivas	Ansioso	Media
Alta	Positivas	Tranquilo	Poca
Alta	Positivas y Negativas	Ansioso	Alta
Alta	Positivas y Negativas	Tranquilo	Media
Baja	Negativas	Ansioso	Media
Baja	Negativas	Tranquilo	Media
Baja	Positivas	Ansioso	Media
Baja	Positivas	Tranquilo	Poca
Baja	Positivas y Negativas	Ansioso	Alta
Baja	Positivas y Negativas	Tranquilo	Media

6.2. Análisis exploratorio

El 64 % (152) de los estudiantes obtuvo una calificación final del curso con valores en el intervalo (3-4), para el 29 % (69) su calificación fue mayor a 4 y el resto obtuvo notas menores a 3.0.

El 59 % de los estudiantes incluidos en el estudio provienen de colegios privados. La distribución de los estudiantes en las facultades de Ingeniería y ciencias de salud fue bastante parecida (24 % y 22 %), se cuenta con mayor participación de estudiantes de la facultad de ciencias administrativas y económicas,y poca participación de estudiantes de la facultad de ciencias de la educación. Se observó también que 110 estudiantes quedaron en la categoría de poca preocupación (46.42 %), 81 en la categoría de preocupación moderada (34.18 %) y 46 en la categoría de alta preocupación (19.40 %). El 66.25 % de los estudiantes tenían que cursar asignaturas (Teoría de probabilidades y Bioestadística) en carreras que se consideran con un grado de matemáticas alto.

$$Facultad = \begin{cases} 1 & Ciencias\ administrativas\ y\ economicas \quad |79(33\,\%)| \\ 2 & Ingenieria \quad |57(24\,\%)| \\ 3 & Decrecho\ y\ Ciencias\ sociales \quad |23(10\,\%)| \\ 4 & Ciencias\ naturales \quad |24(10\,\%)| \\ 5 & Ciencias\ de\ la\ salud \quad |51(22\,\%)| \\ 6 & Ciencias\ de\ la\ educacion \quad |3(1{,}0\,\%)| \end{cases}$$

En la tabla **6-2** se muestra la distribución de los estudiantes de acuerdo con la calificación final del curso y la facultad a la que pertenece, al realizar una prueba de independencia chi cuadrado entre ellas se encontró que la calificación final del curso y la facultad no presentan una asociación estadísticamente importante (Valor p igual a 0.0254).

Tabla **6-2**: Tabla cruzada entre la calificación final del curso y la facultad

	Facultad					
Calificación final del curso	***1***	***2***	***3***	***4***	***5***	***6***
Menor a 3	*7 (3 %)*	*2 (1 %)*	*2 (1 %)*	*1 (0.42 %)*	*4 (2 %)*	*0 (0 %)*
Entre 3 y 4	*60 (25 %)*	*39 (16 %)*	*15 (6 %)*	*14 (5.9 %)*	*23 (10 %)*	*1 (0.42 %)*
Más de 4	*12 (5 %)*	*16 (7 %)*	*6 (3 %)*	*9 (4 %)*	*24 (10 %)*	*2 (1 %)*

El (35.4 %) de los estudiantes de colegio privado, obtuvieron una calificación final del

curso entre 3 y 4, tomando como referencia el colegio de procedencia, se evidencia que la distribución según la calificación final del curso es muy similar para colegio privado y público, encontrando que la mayoría de estudiantes presentan una calificación final del curso entre 3 y 4, seguido por la calificaciones mayores a 4, finalmente, calificaciones menor a 3, al realizar una prueba de igualdad de proporciones chi cuadrado se encontró que efectivamente la distribución de las proporciones entre los tipos de colegio es similar (Valor p igual a 0.1003).

Tabla **6-3**: Tabla cruzada entre la calificación final del curso y el colegio

	Clasificación final del curso		
Colegio	***Menor a 3***	***Entre 3 y 4***	***Más de 4***
Privado	*8 (3.4 %)*	*84 (35.4 %)*	*48 (20.25 %)*
Público	*8 (3.4 %)*	*68 (28.7 %)*	*21 (8.86 %)*

La mayoría de los estudiantes con un grado de matemáticas alto en sus carreras obtuvieron calificaciones entre 3 y 4, con menor porcentaje calificaciones menores a 3. En el caso de estudiantes con un grado de matemáticas menor se observa similar composición, al realizar una prueba de igualdad de proporciones chi cuadrado se encontró que efectivamente la distribución de las proporciones entre el grado de matemáticas no es similar (Valor p igual a 0.009392), aunque el resultado de la prueba está muy cercano al valor crítico de 0.01.

Tabla **6-4**: Cruzada entre la calificación final del curso y el grado de matemáticas

	Clasificación final del curso		
Grado de matemáticas	Menor a 3	Entre 3 y 4	Mas de 4
Alto	10 (4,2¸%)	111 (46,8 %)	36 (15,2 %)
Menor	6 (2,5 %)	41 (17,3 %)	33 (13,9 %)

A continuación, se analiza el comportamiento del índice de preocupación de acuerdo con las variables Grado de dificultad del curso, Razones del grado de dificultad y Sentimiento frente al curso que fueron usadas para su construcción.

Tabla **6-5**: Cruzada entre la clasificación del índice de preocupación y el grado de dificultad que el estudiante asocia al curso

	Clasificación del índice de preocupación		
Grado de dificultad del curso	Alta	Media	Poca
Alta	45 (18.99 %)	62 (26.26 %)	80 (33.76 %)
Baja	1 (0.42 %)	19 (8.02 %)	30 (12.66 %)

La tabla **6-5** muestra la distribución de los niveles del índice de preocupación versus el grado de dificultad del curso manifestado por los estudiantes, esta variable tiene un aporte del 10.08 % en la construcción del índice de preocupación. Se encontró que el 78.91 % de los estudiantes participantes indicaron que el curso tendría un grado de dificultad alto. El mayor porcentaje de estudiantes, es decir, 33.76 % quedaron clasificados en un nivel de preocupación poca y un grado de dificultad alta seguida por el 26.16 % los cuales se clasifican por tener un índice de preocupación media y considerar que el curso tendrá un grado de dificultad alto. Ahora bien, el 12.66 % de los estudiantes presentó poca preocupación y consideró que el grado dificultad del curso seria bajo. Al realizar una prueba de independencia chi cuadrado entre ellas se encontró que el grado dificultad que el estudiante considera para el curso y la clasificación del índice de preocupación presentan una asociación significativa (Valor p igual a 0.001759).

Tabla **6-6**: Cruzada entre la clasificación del índice de preocupación y razones del grado de dificultad

	Clasificación del indice de preocupación		
Razones del grado de dificultad	Alta	Media	Poca
Negativas	38 (16.03 %)	41 (17.30 %)	0 (0 %)
Positivas	0 (0 %)	38 (16.03 %)	110 (46.41 %)
Positivas y Negativas	8 (3.38 %)	2 (0.84 %)	0 (0 %)

La tabla **6-6** aparece la distribución de los niveles del índice de preocupación versus las razones que explican el grado de dificultad que manifestó el estudiante, ésta variable tiene un aporte del 43.86 % en la construcción del índice de preocupación. Se encuentra que los estudiantes que manifiestan tener razones negativas (“Me parecen difíciles las materias que

utilizan las matemáticas", "No tengo el suficiente tiempo para dedicarle a la materia" y "No he visto nada de estadística anteriormente") ninguno fue clasificado con un índice de poca preocupación, en cercanos porcentajes fueron clasificados en nivel de preocupación alto y medio, mientras que los estudiantes que manifestaron tener razones positivas ("Tengo buenos conocimientos en matemáticas", "Estoy interesado y tengo una buena actitud hacia la carrera", "Tengo confianza porque tuve buenos resultados en los cursos previos") quedaron clasificados según el índice construido con niveles de preocupación medio y poco, en mayor porcentaje en este último grupo (46.41 %). Muy pocos estudiantes (4.22 %) indicaron tener razones positivas y negativas para explicar el grado de dificultad que manifestaron. Al realizar una prueba de independencia chi cuadrado entre las razones y el nivel de preocupación se econtró que las razones que explican el grado dificultad que el estudiante considera para el curso y la clasificación del grado de preocupación presentan una asociación estadística significativa (Valor p menor a $2.2*10^{-16}$).

Tabla **6-7**: Cruzada entre la clasificación del índice de preocupación y el sentimiento frente al curso

	Clasifiación del índice de preocupación		
Sentimiento frente al curso	Alta	Media	Poca
Ansioso	46 (19.41 %)	46 (19.41 %)	0 (0 %)
Tranquilo	0 (0 %)	35 (14.77 %)	110 (46.41)

La tabla **6-7** muestra la distribución de los niveles del índice de preocupación versus el sentimiento frente al curso que manifestó el estudiante, está variable tiene un aporte del 46.05 % en la construcción del índice de preocupación. Se encontró que el 46.41 % de los estudiantes indican sentirse tranquilos frente al curso y presenta un nivel de preocupación poca mientras que el 19.41 % que manifestaron sentirse ansiosos presentaron un índice de preocupación alto. Al realizar una prueba de independencia chi cuadrado entre las variables se encontró que las razones que explican el sentimiento frente al curso que el estudiante manifiesta y la clasificación del índice de preocupación presentan una asociación significativa (Valor p menor a $2.2*10^{-16}$).

Al estudiar el comportamiento del índice de preocupación de acuerdo con las variables Calificación final del curso,tipo de Colegio, Grado de matemáticas y Facultad. Se observó para los estudiantes con una calificación menor a 3 preocupación moderada el 25 % presentó

un índice de preocupación alto, el 43.75 % presentó un índice de preocupación moderada y 31.25 % un índice de poca preocupación. En el caso de los estudiantes con calificaciones finales de entre 3 y 4, y más de 4 se evidencia una distribución muy similar, teniendo mayor porcentaje con un índice de poca preocupación . Se realizó una prueba de homogeneidad de proporciones (Chi cuadrado) entre los grupos de calificación y se encontró que hay diferencias estadísticamente significativas en la distribución de dichas proporciones según el índice de preocupación (Valor p es 0.03004).

Tabla **6-8**: Cruzada entre la clasificación del índice de preocupación y la calificación final del curso

	Clasificaciòn del indice de preocupación		
Clasificaciòn final del curso	Alta	Media	Poca
Menor a 3	4 (1.7 %)	7 (3.0 %)	5 (2.1 %)
Entre 3 y 4	37 (15.6 %)	49 (20.7 %)	66 (27.8 %)
Mas de 4	5 (2.1 %)	25 (10.5 %)	39 (16.5 %)

En la tabla **6-9** se encuentra la distribución de estudiantes teniendo en cuenta la clasificación del índice de preocupación y el colegio de procedencia del estudiante, se observó que el 29 % de los estudiantes provienen de colegio privado y presentaron un índice de preocupación poco, lo cual representa para el grupo de estudiantes de colegio privado el 49 % . Para el grupo de estudiantes de colegio público se tiene un mayor porcentaje de estudiantes (42 %) con un índice de preocupación poco seguido por los estudiantes con un índice de preocupación media y finalmente en menor porcentaje estudiantes con índice de preocupación alto (Ver tabla **6-5**).La misma distribución fue observada para los estudiantes provenientes de colegio privado. Se realizó una prueba de homogeneidad de proporciones (Chi cuadrado) entre los grupos de calificación y no se observan diferencias estadísticamente importantes (Valor p es 0.5049).

Tabla **6-9**: Distribución del índice de preocupación entre tipo de colegio

	Clasificación del índice de preocupación		
Colegio	Alta	Media	Poca
Privado	27 (11 %)	44 (19 %)	69 (29 %)
Público	19 (8,0 %)	37 (16 %)	41 (17 %)

En la tabla **6-10** aparece la distribución de los estudiantes teniendo en cuenta la clasificación

del índice de preocupación y el nivel de matemáticas en la carrera. Al realizar una prueba de independencia chi cuadrado entre ellas se encontró que no presentan una asociación significativa (Valor p 0.1618). En la tabla **6-10** se muestra la distribución condicional al grado de matemáticas en carrera; para el grupo con un alto grado de matemáticas en la carrera se tiene un mayor porcentaje de estudiantes (28 %) con un índice de preocupación poca seguido por los estudiantes con un índice de preocupación media y finalmente en menor porcentaje estudiantes con índice de preocupación alto, similar distribución se encontró para el grupo de estudiantes con un menor grado de matemáticas.

Tabla **6-10**: Cruzada entre la clasificación del índice de preocupación y el grado de matemáticas en la carrera

Grado de matematicas en la carrera	Clasificación del índice de preocupación		
	Alta	Media	Poca
Alto	30 (13 %)	60 (25 %)	67 (28 %)
Menor	16 (7,0 %)	21 (9 %)	43 (18 %)

En la tabla **6-11** se encuentra la distribución de estudiantes teniendo en cuenta la clasificación del índice de preocupación y la facultad que se encuentra matriculado el estudiante, se observa baja participación por parte de los estudiantes de la facultad de ciencias de la educación, los estudiantes de esta facultad en su totalidad fueron clasificados con índice de preocupación poca. Se evidencia un mayor porcentaje de estudiantes con un índice de preocupación poca seguido por los estudiantes con un índice de preocupación media y finalmente en menor porcentaje estudiantes con índice de preocupación alto en las restantes facultades.

Tabla **6-11**: Cruzada entre la clasificación del índice de preocupación y la facultad

Facultad	Clasificación del índice de preocupación		
	Alta	Media	Poca
Ciencias Administrativas y Económicas	19 (8.02 %)	26 (10,97 %)	34 (14.35 %)
Ingeniería	6 (2.53 %)	27 (11.39 %)	24 (10.13 %)
Derecho y Ciencias sociales	8 (3.38 %)	6 (2.53 %)	9 (3.80 %)
Ciencias Naturales	7 (2.95 %)	8 (3.38 %)	9 (3.80 %)
Ciencias de la Salud	6 (2.53 %)	14 (5.91 %)	31 (13.08 %)
Ciencias de la Educación	0 (0.0 %)	0 (0.0 %)	3 (1.27 %)

6.3. Modelo de regresión logístico ordinal

En la construcción del modelo de regresión ordinal se inició la integración de las variables regresoras Colegio (X_1), Facultad ($X_{21},\ X_{22},\ X_{23},\ X_{24},\ X_{25}$), Índice de preocupación ($X_{31},\ X_{32}$) y Grado de matemáticas (X_4) en la carrera matriculada ya que se considera podrían afectar la probabilidad de que un estudiante en particular pertenezca a una de las categorías. Se estimaron los modelos y se encontró que:

Tabla **6-12**: Resultados de la estimación de los coeficientes del modelo

Coeficientes	**Estimación**	**valor p**
Colegio	-0.482117	0.0079
Facultad de ingeniería	0.2345040	**0.3073**
Facultad de derecho y C.S	-0.2762415	**0.7422**
Facultad ciencias naturales	0.8194474	0.0095
Facultad de ciencias de la salud	0.1193206	**0.8858**
Facultad ciencias de educación	0.8482721	**0.5143**
Índice de preocupación medio	0.7464743	0.0047
Índice de preocupación bajo	0.7694875	0.0023
α_1	-1.9737884	0.0000
α_2	1.0870852	0.0000

En la tabla **6-12** se observa que la variable Facultad, expresada a través de las variables $X_{21},\ X_{22},\ X_{23},\ X_{24},\ X_{25}$ no presentan un aporte significativo para el modelo planteado, esto podría presentarse debido al hecho de que la variable Grado de dificultad en matemáticas en la carrera fue construida considerando el curso de estadística que tenía matriculado lo cual está ligado directamente con la facultad de donde proviene el estudiante. Debido a este resultado se procedió a estimar el modelo sin incluir la variable Facultad, los resultados fueron, los siguientes:

Tabla **6-13**: Resultado de la estimación de modelo con variables Colegio, índice de preocupación y Grado de matemáticas en la carrera

Coeficientes	**Estimación**	**valor p**
Colegio	-0.4049649	0.0189
Índice de preocupación media	0.7695993	0.0022
Índice de preocupación poca	0.8036286	0.0007
Grado de matemáticas	0.5903073	0.0019
α_1	-2.0978031	0.0000
α_2	0.9146846	0.0001

En la tabla **6-13** se observa que las variables incluidas en el modelo tienen un aporte significativo en la explicación de la probabilidad de que un estudiante sea clasificado en cada una de las tres categorías propuestas (Menor a 3, Entre 3 y 4, Más de 4). Se calculó el R^2 de McFadden como una medida para verificar la calidad del ajuste del modelo, encontrando que sólo se logra explicar el 6.55 % de variabilidad esperada. A continuación, se realiza la prueba de hipótesis de líneas paralelas (El cual indica que lo único que cambia del modelo para categoría es $\alpha_1,\ \alpha_2$), necesaria para validar el supuesto de regresión ordinal. Las hipótesis son:

H_0: Los β_i son iguales para cada nivel de la calificación final del curso

H_1: Los β_i no son iguales para cada nivel de la calificación final del curso.

Para dicha prueba (Prueba de Brant) se encuentra un valor p de 0.10426714, que por ser mayor que 0.01, indica que el procedimiento del modelo de regresión ordinal es viable, ya que no se rechaza la igualdad en los valores de las pendientes.

Ahora bien, se calculó la tasa de clasificaciones correctas, esto es el 64.14 % de los estudiantes evaluados, de los cuales el 85.65 % fueron clasificados en la categoría de calificación entre 3 y 4 mientras que los restantes se clasificaron en la categoría más de 4. En la tabla de clasificación (Tabla **6-14**), se observa que según el modelo ajustado ningún estudiante obtuvo un calificación menor a 3, sin embargo, se sabe que se observaron 16 estudiantes con esta categoría, dichos estudiantes fueron clasificados por el modelo en la categoría entre 3 y 4.

Tabla **6-14**: De categoría de calificación observada versus categoría de calificación del modelo

	Categoría de clasificación		
Categoría de calificación modelo	Menor a 3	Entre 3 y 4	Mas de 4
Menor a 3	0 (0 %)	0 (0 %)	0 (0 %)
Entre 3 y 4	16 (6.75 %)	135 (56.96 %)	52 (21.94 %)
Mas de 4	0 (0 %)	17 (7.17 %)	17 (7.17 %)

A pesar de que el modelo cumple con el supuesto de líneas paralelas, la tasa de clasificación es baja y el R^2 de McFadden es muy bajo. Con base en lo expuesto se decidió usar el modelo como medida descriptiva para identificar las variables que de forma conjunta puede afectar

la probabilidad de un estudiante pertenecer a una de las categorías con las que se realizará el modelo bayesiano.

6.4. Modelo bayesiano

En esta sección se toma como referencia, los grupos de combinaciones de variables formadas por las opciones de respuestas de las variables: Colegio (X_1), índice de preocupación ($X_{31},\ X_{32}$) y Grado de matemáticas (X_4), variables que resultaron significativas en el modelo de regresión ordinal estimado, para explicar el comportamiento de la variable calificación final del curso. A continuación, se muestran los grupos conformados por:

Tabla **6-15**: Conformación de grupos de estudiantes

Grupo	Colegio	Indice de preocupación	Nivel de matematicas
1	Privado	Alto	Alto
2	Privado	Moderado	Alto
3	Privado	Poco	Alto
4	Publico	Alto	Alto
5	Publico	Moderado	Alto
6	Publico	Poco	Alto
7	Privado	Alto	Menor
8	Privado	Moderado	Menor
9	Privado	Poco	Menor
10	Publico	Alto	Menor
11	Publico	Moderado	Menor
12	Publico	Poco	Menor

De acuerdo con la tabla **6-16**, se observa que en algunos grupos como 8, 9, 10 (Tienen en común que el grado de matemáticas de sus carreras es menor) no presentaron estudiantes con calificaciones finales menores a 3 mientras que el grupo 1 (Estudiantes provenientes de colegio privado, con un índice de preocupación alto y un grado de matemáticas en su carrera alto) no presentó estudiantes con calificaciones finales superiores a 4 (Sólo 3 estudiantes con calificaciones finales menores a 3), ahora bien, comparándolo con el grupo 4 (La única diferencia es la procedencia del estudiante de colegio público) se observa que todos los estudiantes obtuvieron calificaciones entre 3 y 4. Cabe resaltar que en los grupos 7, 8 y 12 la mayoría de los estudiantes obtuvo una calificación mayor a 4.

Tabla **6-16**: Cantidad de estudiantes clasificados por grupo y calificación final

Grupo	Menor a 3	Entre 3 y 4	Mas de 4
1	3	16	0
2	2	22	11
3	2	26	16
4	0	11	0
5	2	17	6
6	1	19	3
7	1	3	4
8	0	3	6
9	0	14	11
10	0	7	1
11	3	7	2
12	2	7	9

Tomando en cuenta esta información y las distribuciones a priori no informativas sugeridas por Perks (1947) y Jeffreys (1946), se calculó la probabilidad predictiva de que un nuevo estudiante obtenga una calificación final en alguna de las tres categorías para cada grupo conformado (Tabla **6-17**). Se observó que las probabilidades predictivas usando ambas distribuciones a priori no presentan muchas diferencias para cada grupo. Para la mayoría de los grupos, exceptuando los grupos 7, 8 y 12, la calificación final del curso más probable es entre 3 y 4. En los grupos 7, 8 y 12 la calificación final del curso con mayor probabilidad es más de 4, la característica que comparten es un grado menor de matemáticas de sus carreras. En general, se encuentra que para un estudiante con un grado alto de matemáticas en la carrera para cualquier combinación de las restantes condiciones (Colegio e índice de preocupación) es mayormente probable que el estudiante obtenga una calificación final entre 3 y 4. Se debe resaltar que en los grupos 1, 4, 6 y 10 la probabilidad de que un nuevo estudiante con estas condiciones obtenga una calificación entre 3 y 4 es superior a 0.8 (Posibilidad superior al 80 %).

Tabla **6-17**: Probabilidades predictivas a posteriori para cada grupo de estudiantes

	Perks (1947)			Jeffreys (1946)		
Grupo	Menor a 3	Entre 3 y 4	Mas de 4	Menor a 3	Entre 3 y 4	Mas de 4
1	0.17	0.82	0.02	0.17	**0.80**	0.02
2	0.06	0.62	0.31	0.07	0.62	0.32
3	0.05	0.59	0.36	0.05	0.58	0.36
4	0.03	0.94	0.03	0.04	**0.92**	0.04
5	0.09	0.67	0.24	0.09	0.66	0.25
6	0.06	0.81	0.14	0.06	**0.80**	0.14
7	0.15	0.37	0.48	0.16	0.37	0.47
8	0.03	0.33	0.63	0.05	0.33	**0.62**
9	0.01	0.55	0.44	0.02	0.55	0.43
10	0.04	0.81	0.15	0.05	**0.79**	0.16
11	0.26	0.56	0.18	**0.26**	0.58	0.19
12	0.12	0.39	0.49	0.13	0.38	0.49

7 Conclusiones

- El modelamiento del desempeño de los estudiantes en un primer curso de estadística, es viable de acuerdo con el instrumento diseñado para recolectar la información por(Tovar Cuevas y cols., 2010), y el cual ha sido modificado para adaptarlo a la institución donde se aplicó la encuesta.

- El índice de preocupación o ansiedad con el que los estudiantes llegan al curso, se puede construir a partir de metodología estadística multivariada. Su nivel de clasificación para la información analizada es satisfactorio y adecuado.

- El modelo de respuesta ordinal, inicialmente propuesto en los objetivos, para el modelamiento del desempeño de los estudiantes en su primer curso de estadística, no logró la bondad de ajuste esperada, sin embargo, sirvió de manera descriptiva para seleccionar las variables que más influirían en el desempeño final de los estudiantes, para incluirlas en un modelo con metodología bayesiana.

- El poder de clasificación y predicción del modelo bayesiano desarrollado para describir el desempeño de los estudiantes, cumple con todos los requisitos estadísticos para clasificarlo como un buen modelo matemático y el índice de preocupación resulta ser una variable influyente en dicho modelo.

8 Anexos

```
##Análisis exploratorio.

table(Cod_nota_final); table(Colegio); table(Facultad); table(GradDifMate)
#Prueba chi-cuadrado
chisq.test(Facultad,GradDifMate)

##Ajuste del modelo de regresión logístico ordinal.

library(MASS)
Cod_nota_21<-ordered(Cod_nota_final,levels=c("<3","[3-4)","[4-5]"))
m <- polr(Cod_nota_21~as.factor(Indice.Clasificado)+as.factor(Colegio)+
as.factor(GradDifMate), Hess=TRUE,method="cloglog")
## Cálculo de los valor p
p <- pnorm(abs(ctable[, "t value"]), lower.tail = FALSE) * 2

##Pruebas de lineas paralelas

library(brant)
brant(m,by.var=TRUE)
##R2 de MCFadden
library(pscl)
pR2(m)\\

###Modelo bayesiano

Base<-read.table("clipboard",header=TRUE,sep="\t",dec=",",row.names=1)
attach(Base)
```

```
##Número de grupo
grupo=1
library(R2OpenBUGS)
sink("modelo.txt")
cat("
    model{
    y[1]<-l
    y[2]<-m
    y[3]<-ñ
    y[1:3]~dmulti(phi[1:3],n)
    phi[1:3]~ddirch(alpha[1:3])
    alpha[1]<-a
    alpha[2]<-b
    alpha[3]<-c
    }", fill=TRUE)
sink()

datos_jefrey=list(a=0.5,b=0.5,c=0.5,n=Total.general[grupo],l=X.3[grupo],m=X.3.4.
[grupo],ñ=X.4.5.[grupo])
datos_no_informativa=list(a=1/3,b=1/3,c=1/3,n=Total.general[grupo],
l=X.3[grupo],m=X.3.4.[grupo],ñ=X.4.5.[grupo])parametros=list("phi")

mod_jefrey=bugs(data=datos_jefrey,inits=NULL,parameters.to.save = parametros,
model.file = "modelo.txt",n.iter=10000,n.burnin = 5000,n.thin = 1,n.chains = 1,
codaPkg = FALSE,debug = FALSE)

mod_no_informativa=bugs(data=datos_no_informativa,inits=NULL,parameters.to.save
= parametros,model.file = "modelo.txt", n.iter=10000,n.burnin = 5000,n.thin
= 1,n.chains = 1,codaPkg = FALSE,debug = FALSE)
```

Referencias

Avella, B., Oliva, M., y cols. (2010). *Comparación del análisis factorial múltiple (afm) y del análisis en componentes principales para datos cualitativos (prinqual), en la construcción de índices/comparison between multiple factor analysis (mfa) and principal component analysis for qualitative data (prinqual) methods for derivation of indices* (Tesis Doctoral no publicada). Universidad Nacional de Colombia.

Batanero, C., Godino, J. D., Vallecillos, A., Green, D. e., y Holmes, P. (1994). Errors and difficulties in understanding elementary statistical concepts. *International Journal of Mathematical Education in Science and Technology*, *25*(4), 527–547.

Behar, R., Ojeda Ramírez, M. M., y cols. (2011). El problema de la educación estadística: perspectivas desde el aprendizaje.

Burt, C. (1950). The factorial analysis of qualitative data. *British Journal of Statistical Psychology*, *3*(3), 166–185.

Carmona Márquez, J. (2004). Una revisión de las evidencias de fiabilidad y validez de los cuestionarios de actitudes y ansiedad hacia la estadística. *Statistics Education Research Journal*, *3*(1), 5–28.

Castaño Vélez, E. A., y cols. (2010). Estimación del indicador de calidad de vida para el departamento de antioquia.

Cheng, J., Wang, Z., y Pollastri, G. (2008). A neural network approach to ordinal regression. En *2008 ieee international joint conference on neural networks (ieee world congress on computational intelligence)* (pp. 1279–1284).

Chevallard, Y. (2007). *Passé et présent de la théorie anthropologique du didactique.*

Connor, R., y Mosimann, J. (s.f.). Conceptos de independencia para proporciones con una generalización de la distribución de dirichlet. *Diario de la Asociación Americana de Estadística*.

Dobson, A. (2012). *An introduction to generalized linear models.* Editorial Wiley Sons.

Fardales Macias, V. E., Diéguez Batista, R., y Puga García, A. (2012). Tendencias históricas del proceso de formación estadística del profesional de medicina. *Gaceta Médica Espirituana*, *14*(2), 157–166.

Freud, J., Miller, I., y M, M. (2000). Estimación bayesiana. En (p. 353). Pearson Educación. (ISNB: 9789701703892)

Gameroff, M. J. (2005). Using the proportional odds model for health-related outcomes: Why, when, and how with various sas® procedures. En *Sugi* (Vol. 30, pp. 205–230).

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., y Rubin, D. B. (1995). Bayesian data analysis. En (Vol. 2, p. 8). CRC press Boca Raton, FL.

Gil Flores, J. (1999). Actitudes hacia la estadística. incidencia de las variables sexo y formación previa. *Revista española de pedagogía*, 567–589.

Guttman, L. (1941). The quantification of a class of attributes: A theory and method of scale construction. *The prediction of personal adjustment*, 319–348.

Hayashi, C. (1956). Theory and examples of quantification.(ii). En *Proc. of the institute of statist. math* (Vol. 4, pp. 19–30).

Lebart, L., Morineau, A., y Piron, M. (1995). *Statistique exploratoire multidimensionnelle* (Vol. 3). Dunod Paris.

Minotta, E. (2018). *Factores propios de la educación media y su relación con el desempeño en el primer curso de estadística en la universidad.* (Trabajo de grado para optar al título de Estadístico)

Osorno, S. (2015). *Preconceptos de los estudiantes sobre el primer curso de estadística en la universidad.* (Trabajo de grado para optar al título de Estadístico)

Pardo, C. E., y Cabarcas, G. (2001). *Métodos estadísticos multivariados en investigación social.*

Perks, W. (1947). Some observations on inverse probability including a new indifference rule. *Journal of the Institute of Actuaries*, *73*(2), 285–334.

Ponteville, C. C. (2014). ¿ para qué enseñamos estadística?

Roberts, D. M., y Saxe, J. E. (1982). Validity of a statistics attitude survey: A follow-up study. *Educational and psychological measurement*, *42*(3), 907–912.

Rodriguez, M. Á. D. (2008). Modelos de respuesta discreta en r y aplicación con datos reales. *Universidad de Granada*.

Salazar, L., y Jairo, J. (2017). Enseñando estadística a futuros ingenieros. *Scientia et technica*, *1*(34).

Tovar Cuevas, J. R. (2015). Inferencia bayesiana e investigación en salud: un caso de aplicación en diagnóstico clínico. *Revista Médica de Risaralda*, *21*(1).

Tovar Cuevas, J. R., Castillo, H., y Marín, M. d. P. (2010). Preconcepciones de estudiantes de la pontificia universidad javeriana cali sobre el curso de estadística. *Pensamiento Psicológico*, *3*(9).

Yay, M., y Akıncı, E. D. (2009). Application of ordinal logistic regression and artificial neural networks in a study of student satistaction. *Cypriot Journal of Educational Sciences*, *4*(7), 58–69.

Resumen

Los estudiantes universitarios que ingresan a un primer curso de estadística, llegan con preconcepciones, temores, y con algunos conocimientos básicos, si han tenido la oportunidad en los colegios de secundaria, de ver algunos temas relacionados con la estadística. En muchas de las universidades de carácter privado, el fracaso de los estudiantes, originado por afrontar los primeros cursos de matemática universitaria y en este caso en particular, por afrontar un primer curso de estadística, es de gran preocupación. Es en este sentido que la presente investigación quiere indagar, especialmente en alguna de las universidades privadas del sur de Cali, los conocimientos previos, antecedentes académicos, temores, entre otros factores que podrían amenazar el rendimiento y resultado de los estudiantes en un primer curso universitario de estadística, teniendo en cuenta carreras con fuerte componente matemático y las que no lo tienen.

Palabras claves: Desempeño académico, Modelo de regresión ordinal, Estadística Bayesiana, Probabilidad predictiva