Capítulo 2

2. Marco Teórico y Definición de Variables

2.1. Introducción

El propósito de este capítulo es proporcionar un conocimiento básico en estadística descriptiva, inferencial y multivariada; teniendo en cuenta las necesidades del lector para poder comprender los análisis que se presentarán más adelante.

Además se incluye la descripción de cada una de las variables que involucran este trabajo de manera que sea fácilmente entendible su uso y el respectivo análisis.

2.2. ¿Qué es la Estadística?

Es el sistema que estudia los métodos científicos para recopilar datos, analizarlos, agruparlos y presentarlos adecuadamente para su interpretación.

2.3. Campo de aplicación de la estadística

La teoría general de la estadística es aplicable a cualquier campo científico en el cual se hacen observaciones. La necesidad del hombre moderno por utilizar la estadística es cada vez más amplia y profunda.

Hasta hace algunos años, el estudio de esta materia aparece sólo en algunos programas de nivel universitario; actualmente se considera como una disciplina esencial en todos los campos de la investigación.

El análisis estadístico es la aplicación de técnicas que permiten hacer inferencias sobre datos correspondientes a una población, a través de procedimientos específicos; entre las técnicas que se pueden aplicar a la investigación estadística se encuentra el análisis univariado y multivariado, ambos permiten interpretar valores con resultados primordiales al servicio de la sociedad.

2.4. Definiciones Básicas

- **Experimento:** Es un proceso por medio del cual se obtiene una observación o medida cualquiera.

- **Campo de Borel:** Toda familia de conjuntos abiertos que recubren un conjunto cerrado y acotado contiene una subfamilia finita que lo recubre. El campo de Borel es aquel conjunto que contiene todas las uniones contables de dichos subconjuntos.

- **Mínimo Campo de Borel:** Es la intersección de todos los campos de Borel de dicho conjunto.

- **Espacio muestral:** Sea ? el conjunto de todos los resultados posibles de un experimento, sea $ el mínimo campo boreal de ?, el par **(? ,**$) se denomina espacio muestral.

- **Variable aleatoria:** Sea (? ,$) un espacio muestral con una medida de probabilidad y **x** es una función de valor real definida con respecto a los elementos de (? ,$), entonces **x** se denomina variable aleatoria.

- **Parámetro poblacional:** Es el valor que caracteriza a una población y está denotado por ?.

- **Estimador:** Dada una muestra de tamaño n, el estimador de una población ? es una función $\hat{q} : R^n \rightarrow R$ tal que en su definición no está incluido el valor de ?.

- **Variable aleatoria discreta:** Se dice que una variable aleatoria es discreta si y sólo si el número de valores que puede tomar es finito o numerable.

  En este caso la función de probabilidad de x se denota por f(x)=P(X=x), tal que:

  i) $0 \leq f(x) \leq 1$

  ii) $\sum_x f(x) = 1$

  La función de distribución acumulada de la variable aleatoria discreta x, está definida por:

  $$F(x) = P(X \leq x) = \sum_{t \leq x} f(t)$$

  donde f(t) es el valor de la distribución de probabilidad de x en t.

- **Variable aleatoria continua:** Se dice que **x** es una variable continua si sus valores consisten en uno o más intervalos de la recta real.

En este caso la función de probabilidad de x cumple con las siguientes condiciones:

i) $f(x) \geq 0$, $\forall x \in D_f$

ii) $\int_{-\infty}^{\infty} f(x)dx = 1$

iii) $p(a \leq X \leq b) = \int_{b}^{a} f(x)dx$

La función dada por $F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt$ para $-\infty < x < \infty$

donde f(t) es el valor de la distribución de probabilidad de x en t.

- **Valor esperado de una variable aleatoria:** Sea f(x) la distribución de probabilidad de una variable aleatoria discreta x, el valor esperado de esta variable aleatoria g(x) es:

$E[g(x)] = \sum_{x} g(x) \cdot f(x)$

Si g(x) = x $\rightarrow E(x) = \sum_{x} x \cdot f(x)$

Si g(x) = $(x-\mu)^2$ $\rightarrow E[g(x)] = \sum_{x} (x-m)^2 \cdot f(x)$

El valor esperado da función g(x) = x se conoce como la media poblacional $\overline{C}$, y el valor esperado da función g(x) = $(x-\mu)^2$ se conoce como la varianza poblacional $\mathbf{S^2}$.

Sea f(x) la función de densidad de una variable aleatoria continua x, el valor esperado de esta variable aleatoria g(x) es:

$$E[g(x)] = \int_{-\infty}^{\infty} g(x) \cdot f(x) dx$$

Si g(x) = x $\rightarrow E(x) = \int_{-\infty}^{\infty} x \cdot f(x) dx$

Si g(x) = $(x-\mu)^2$ $\rightarrow E[g(x)] = \int_{-\infty}^{\infty} (x-m)^2 \cdot f(x) dx$

2.4.1. Estadística descriptiva

La estadística descriptiva es un conjunto de técnicas que permiten la interpretación y análisis de los datos de una población. Entre estas técnicas tenemos las tablas y gráficos de frecuencia absoluta y relativa, y las medidas de tendencia central, dispersión, sesgo y coeficiente de Kurtosis de una muestra.

- **Mediana poblacional:** Dada una muestra de tamaño n, la mediana es el valor central dentro de ese conjunto de datos, es decir que la mitad de los datos están por arriba de este valor y la otra mitad está por debajo de el.

  La mediana es el $\left(\frac{n+1}{2}\right)$ ésimo término del arreglo de datos. Si el valor de n es par, la mediana será el promedio entre el $\left(\frac{n}{2}\right)$ ésimo término y el $\left(\frac{n+2}{2}\right)$ ésimo término.

- **Moda Poblacional:** Dada una muestra de tamaño n, la moda es el valor que más se repite dentro de ese conjunto de datos.

- **Desviación estándar poblacional:** Este estimador mide la variabilidad de las observaciones con respecto a la media poblacional y es la raíz cuadrada de la varianza poblacional.

- **Coeficiente de sesgo:** Una distribución es simétrica o insesgada si los valores de la media, la mediana y la moda son iguales; caso contrario la distribución es asimétrica.

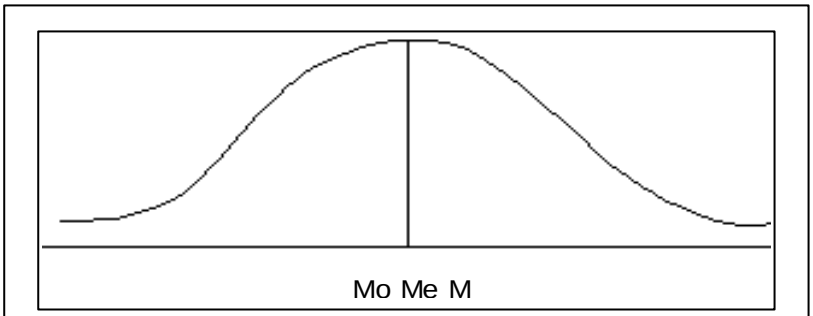

Gráfico 2.1. Distribución Simétrica

Cuando existe asimetría el valor de la moda no se produce cambio en la moda, pero la mediana y la media se corren en dirección de la asimetría.

Para determinar el coeficiente de sesgo se calcula el tercer momento con respecto a la media m_3, definido de la siguiente manera:

$$a_3 = \frac{m_3}{s^3}$$

En la asimetría positiva hacia la derecha la mediana aumenta por el mayor número de frecuencias hacia la derecha y la media aumenta más, ya que hay un incremento en la frecuencia y en el valor de las observaciones, en este caso se dice que el sesgo es positivo.

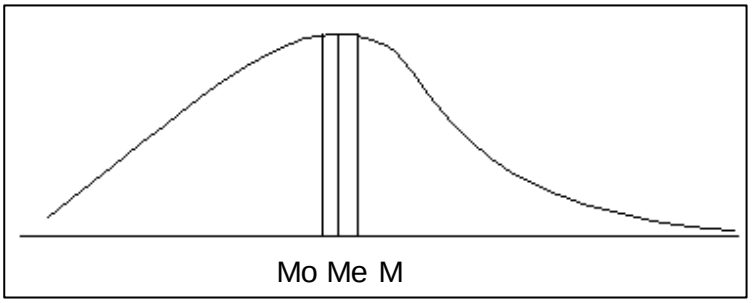

Gráfico 2.2. Distribución Asimétrica Positiva

En la asimetría negativa ocurre lo contrario; la mediana disminuye y la media se reduce más que la mediana, entonces el coeficiente de sesgo es negativo

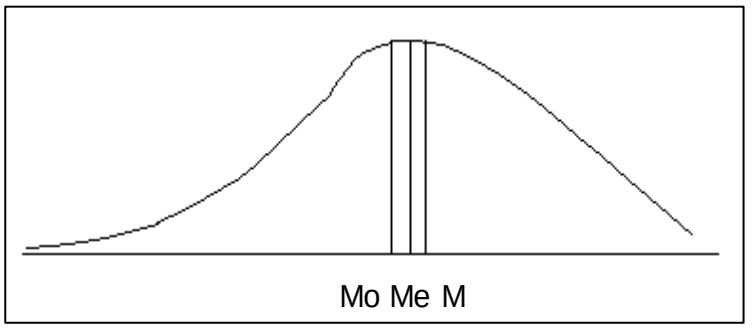

Gráfico 2.3. Distribución Asimétrica Negativa

- **Coeficiente de Kurtosis:** Permite apreciar el grado en que una curva de distribución de frecuencias es más alta o más achatada que la curva normal de distribución.

Para determinar el coeficiente de kurtosis se calcula el cuarto momento con respecto a la media m_4, definido de la siguiente manera:

$$a_4 = \frac{m_4}{s^4}$$

Para la curva de distribución normal $a_4 = 3$. Si a >3, la curva es leptocúrtica y si a < 3, la curva es platicúrtica.

2.4.2. Estadística inferencial

La estadística inferencial es un conjunto de métodos utilizados para tomar decisiones o para obtener conclusiones de una población. El presente trabajo hará uso de pruebas de hipótesis para obtener inferencias estadísticas

- **Hipótesis estadística:** Es una afirmación acerca de los parámetros de una población o a la distribución de tal población o ambos.

H0 es la hipótesis nula, la cual usualmente se desea rechazar.

H1 es la hipótesis alterna.

Se llama prueba estadística al procedimiento que permite rechazar o no la hipótesis nula, el cual está basada en la información que proporciona una muestra aleatoria.

Al rechazo de la hipótesis estadística cuando ésta es verdadera se la conoce como *error de tipo I*, la probabilidad de cometer un error de tipo I se denota por a.

$a = P(\text{ rechazar } H_0 / H_0)$

Cuando la hipótesis nula es aceptada siendo falsa se denomina *error tipo II*; la probabilidad de cometer un error de tipo II se denota por ß.

$ß = P(\text{ rechazar } H_1 / H_1)$

A la región de rechazo de H_0 se la denomina región crítica de la prueba; el tamaño de la región crítica es la probabilidad de cometer un error tipo I, el cual también se denomina nivel de significancia de la prueba.

2.5. Técnicas Multivariado

En el análisis estadístico se puede obtener conclusiones significativas al hacer uso de las herramientas multivariadas; el presente estudio utilizará esas herramientas para determinar si existe algún efecto en la interacción de las variables, para esto se necesita tener claro algunas definiciones importantes que se detallarán a continuación.

- **Matriz de datos:** Supóngase la existencia de n unidades de investigación y p características investigadas, con estos datos se construirá una matriz de datos que estará formada por n filas y p columnas.

Tabla 2.1

Matriz de datos

		Variables			
		Variable 1	**Variable 2**	**...**	**Variable p**
Unidades de investigación	**1**	X_{11}	X_{12}	…	X_{1p}
	2	X_{21}	X_{22}	…	X_{2p}
	⋮	:	:	:	:
	n	X_{n1}	X_{n2}	…	X_{np}

Luego de haber ordenado la matriz de datos se procede a obtener la matriz de varianzas y covarianzas **S**, la cual tiene se genera a través de la siguiente fórmula:

$$\hat{m} = \left[\overline{c}_1, \overline{c}_2, ..., \overline{c}_p\right] = \left[\hat{m}_1, \hat{m}_2, ..., \hat{m}_p\right]$$

$$\overline{c}_j = \frac{1}{n}\sum_{i=1}^{n}\left(c_{ij}\right) \qquad \text{para} \qquad \mathbf{j=1,2,\ldots p}$$

$$S_{jk} = \frac{1}{n}\sum_{i=1}^{n}\left(c_{ij} - c_j\right)\left(c_{kj} - c_k\right) = \hat{s}_{jj} \qquad \text{para} \qquad j \neq k$$

$$S_{jj} = \frac{1}{n}\sum_{i=1}^{n}\left(c_{ij} - c_j\right)^2 = \hat{s}_{jk} \qquad \text{para} \qquad j = 1,2,\ldots p$$

2.5.1. Matriz de correlación

Los coeficientes para la matriz de correlación están dados por:

$$R_{ij} = \frac{S_{ij}}{\sqrt{S_{ij}S_{jj}}} = \hat{r}_{ij}$$

$$R_{ij} = R_{ji} \quad \text{para} \quad j \neq i$$

La matriz de correlación, denotada por R, será de p filas y p columnas, la cual está conformada por unos en su diagonal principal.

Tabla 2.1

Matriz de correlación

R =

	Variable 1	**Variable 2**	**…**	**Variable p**
Variable 1	1	R_{12}	…	R_{1p}
Variable 2	R_{21}	1	…	R_{2p}
:	:	:	:	:
Variable p	R_{n1}	R_{n2}	…	1

Se considerará la iteración R_{ij} como alta correlación, si en valor absoluto este valor es mayor que 0.6.

2.5.2. Tablas de contingencia

El método bivariado que permite determinar si existe o no dependencia entre las variables de una muestra o población se denomina “tablas de contingencia”. En el planteamiento de la hipótesis nula se propone que existe independencia entre un par de variables versus la hipótesis alterna de dependencia.

Tabla 2.2

Modelo de una tabla de contingencia

		Variable 2				
Variable 1		**2-1**	**2-2**	**…**	**2-c**	**total**
1-1	F.O. V.E. F.R.O.	f11 e11	f12 e12	… …	f1c e1c	**f1.** **e1.**
1-2	F.O. V.E. F.R.O.	f21 e21	f22 e22		f2c e2c	**f2.** **e2.**
⋮	F.O. V.E. F.R.O.	⋮	⋮			⋮
1-r	F.O. V.E. F.R.O.	fr1 er1	fr2 er2	… …	frc erc	**fr.** **er.**
Total	F.O. V.E. F.R.O.	**f.1** **e.1**	**f.2** **e.2**		**f.c** **e.c**	**f..** **e..**

Donde f_{ij} es la frecuencia observada de la celda de i-ésimo renglón y la j-ésima columna, $f_{i.}$ y $f_{.j}$ los totales de los renglones y columnas respectivamente, y $f_{..}$ el total de la suma de todas las frecuencias de las celdas.

Si $?_{ij}$ es la probabilidad de que un elemento quede en la celda perteneciente al i-ésimo renglón y la j-ésima columna, $?_{i}.$ es la probabilidad de que un elemento quede en el i-ésimo renglón, y $?._{j}$ es la probabilidad de que un elemento quede en la j-ésima columna, la hipótesis nula será $?_{ij} = ?_{i}. * ?._{j}$ para i = 1,2,…,r y

j = 1,2,…,c. La hipótesis alterna entonces sería $?_{ij} \neq ?_{i} . * ?._{j}$ cuando menos para una pareja de valores i y j.

$$\hat{q}_{i.} = \frac{f_{i.}}{f} \qquad \text{y} \qquad \hat{q}_{.j} = \frac{f_{.j}}{f}$$

Y con la hipótesis nula de independencia se tiene

$$e_{ij} = \hat{q}_{i.} \cdot \hat{q}_{.j} \cdot f = \frac{f_{i.}}{f} \cdot \frac{f_{j.}}{f} \cdot f = \frac{f_{i.} \cdot f_{.j}}{f}$$

Donde e_{ij} es la frecuencia esperada de la celda en el i-ésimo renglón y la j-ésima columna. La decisión de aceptar o rechazar la hipótesis nula se basa en la siguiente fórmula:

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(f_{ij} - e_{ij})^2}{e_{ij}}$$

De esta manera se rechaza la hipótesis nula si este valor excede al ji cuadrado $c^2_{a,(r-1)(c-1)}$ con (r-1)(c-1) grados de libertad.

2.5.3. Análisis de componentes principales

El propósito de aplicar componentes principales es obtener un nuevo grupo de k variables, a través de la información que proporcionan las p variables originales, tal que k<p.

Sea $X=(X_1, X_2,...X_p)$ un vector aleatorio p-variado con matriz de media **μ** y matriz de varianzas y covarianzas $\sum$, supóngase además que los valores propios de $\sum$ son $l_1, l_2,...l_p$; definamos p variables no observadas $Y_1,Y_2,...Y_p$, como la combinación lineal de $X_1, X_2,...X_p$, dado por:

$$Y_i = b_{1i}X + b_{2i}X_2 + ...b_{pi}X_p = \begin{bmatrix} b_{1i} & b_{2i}...b_{pi} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix} = b_i^T X$$

Donde $b_i^T = \begin{bmatrix} b_{1i} & b_{2i}...b_{pi} \end{bmatrix}$ para i = 1,2,…p; es el vector propio de la matriz de varianzas y covarianzas $\sum$ de cada una de las p variables observadas.

$$E[Y_i] = E[b_i^T X] = b_i^T E[X] = b_i^T m \qquad \text{i = 1,2,…p}$$

$$\text{var}\, Y_i = b_i^T \sum b_i = l_i$$

$$\text{cov}(Y_i, Y_j) = b_i^T \sum b_j = 0 \qquad i \neq j$$

Las componentes principales de $\mathbf{X}$ son aquellas combinaciones lineales $Y_1, Y_2, ...Y_p$ que son no correlacionadas entre sí y cuyas varianzas sean tan grandes como sea posible.

Lo que se busca es obtener la mayor proporción de la variación de la población explicada por las componentes donde el valor individual de su aporte está dado por $?_k$ / ($?_1$ +$?_2$ +…+$?_p$) para k=1,2,…,p.

El número de componentes principales escogidas dependerá del porcentaje de varianza que se desea explicar, lo que depende del tipo de estudio que se está realizando, (generalmente se trabaja con un mínimo de 80% de la variación explicada por las componentes.)

El siguiente paso es describir si es posible la relación de covarianza de muchas variables en término de unas pocas variables no observables llamadas factores, este estudio recibe el nombre de análisis de factores.

El análisis de factores utiliza la matriz de carga, que es una matriz de correlaciones entre las componentes principales y las variables originales, en la cual se pueden formar factores que estén altamente correlacionadas y así identificar las más representativas de cada componente principal.

Se recomienda usar la matriz de correlación en lugar de la matriz de carga, ya que al estandarizar no puede existir una variable con alta varianza que pueda influir en la determinación de la matriz de carga.

Por último se puede acotar que una rotación de factores, denominada **Varimax** puede ser útil para obtener factores más claros que identifiquen cada componente. La aplicación de Varimáx es una rotación de los factores de forma ortogonal.

2.6. Descripción de variables

En el presente capítulo se definen las variables que se estudiarán para medir el aprovechamiento de los créditos educativos para los alumnos de la ESPOL, detallando cada una de las 13 variables.

Las 5 primeras variables: Carrera del prestatario, Nivel del prestatario, Tiempo del préstamo, rubro y Monto del préstamo fueron obtenidas a través de la base de datos que tiene el departamento de Bienestar Educativo de la ESPOL

Las 8 últimas variables: Número de materias aprobadas en el primer, segundo, tercero y cuarto semestre, y el promedio de materias aprobadas en los mismos semestres se obtuvo a través del sistema académico, el cual maneja el historial educativo de cada uno de los alumnos de la ESPOL.

2.6.1. Variable C_A: Carrera del Prestatario

La variable cualitativa nominal C_A representa la carrera que estaba estudiando el alumno en el momento que realizó el préstamo al IECE.

Esta variable puede tomar 21 valores distintos:

- Economía
- Ing. Acuicultura
- Ing. Agropecuaria
- Ing. Alimentos
- Ing. Comercial
- Ing. Computación
- Ing. Civil

- Ing. Eléctrica - Industrial
- Ing. Estadística - Informática
- Ing. Geología
- Ing. y administración en la producción Industrial
- Ing. Mecánica
- Ing. Naval
- Ing. Eléctrica - Potencia
- Lic. Sistema de información
- Lic. Turismo
- Tecn. Alimentos
- Tecn. Computación
- Tecn. Eléctrica
- Tecn. Mecánica
- Tecn. Pesquería

2.6.2. Variable N_V: Nivel del Prestatario

La variable cualitativa ordinal N_V representa el nivel en el que se encontraba el estudiante en el momento de solicitar su crédito.

Esta variable puede tomar los siguientes valores:

- Nivel 100 – I
- Nivel 100 - II
- Nivel 200 - I
- Nivel 200 – II
- Nivel 300 - I
- Nivel 300 - II
- Nivel 400 - I

- Nivel 400 - II
- Nivel 500 - I
- Nivel 500 - II

2.6.3. Variable T_P: Tiempo del Préstamo

La variable cualitativa ordinal T_P representa el número de semestres para los cuales el estudiante hizó el préstamo al IECE. El tiempo del préstamo puede tomar valores de 2 a 12 semestres.

2.6.4. Variable R_P: Rubros del préstamo

La variable cualitativa nominal R_P representa el rubro o motivo para el cual el estudiante solicitó el crédito del IECE, el cual esencialmente puede ser para registros o pensiones hasta finalizar los estudios, para sostenimiento, este rubro se otorga a estudiantes que viven fuera de Guayaquil, para elementos de estudio, es decir libros, para derechos de grado y para realizar tesis o tópicos.

Esta variable puede incluir combinaciones de los rubros antes mencionados; los valores que puede tomar son:

- Sólo registros
- Sólo sostenimiento
- Sólo elementos de estudio
- Registros y sostenimiento
- Registros y elementos de estudio
- Sostenimiento y elementos de estudio
- Registros, sostenimiento y elementos de estudio
- Registros y tópico
- Registros, elementos de estudio y tópico
- Registros y derechos de grado
- Registros, derechos de grado y elementos de estudio

2.6.5. Variable M_P: Monto del Préstamo

La variable cuantitativa real M_P representa la cantidad en dólares a la que asciende el crédito del estudiante. Esta cantidad va desde $106 hasta $1.398.

2.6.6. Variable N_M_1_S: Número de Materias Aprobadas en el Primer Semestre

La variable cualitativa ordinal N_M_1_S representa el número de materias aprobadas en el primer semestre con crédito educativo. Esta variable puede tomar valores de 1 a 8.

2.6.7. Variable N_M_2_S: Número de Materias Aprobadas en el Segundo Semestre

La variable cualitativa ordinal N_M_2_S representa el número de materias aprobadas en el segundo semestre con crédito educativo. Esta variable puede tomar valores de 1 a 10.

2.7.8. Variable N_M_3_S: Número de Materias Aprobadas en el Tercer Semestre

La variable cualitativa ordinal N_M_3_S representa el número de materias aprobadas en el tercer semestre con crédito educativo. Esta variable puede tomar valores de 1 a 9.

2.7.9. Variable N_M_4_S: Número de Materias Aprobadas en el Cuarto Semestre

La variable cualitativa ordinal N_M_4_S representa el número de materias aprobadas en el cuarto semestre con crédito educativo. Esta variable puede tomar valores de 1 a 10.

Estas cuatro variables sirven para hacer un seguimiento al estudiante desde el momento en el que es beneficiario del crédito del IECE, al igual que las próximas 4 variables que miden el rendimiento educativo.

2.7.10. Variable P_1_S: Nota Promedio de Materias Aprobadas en el Primer Semestre

La variable cuantitativa real P_1_S representa la nota promedio de los estudiantes en el primer semestre con crédito de IECE. Las calificaciones de los estudiantes de la ESPOL pueden ir de 0.00 a 10.00, pero la mínima nota que un estudiante necesita para aprobar una materia es de 6.00, por lo que esta variable puede tomar valores de 6 a 10.

2.7.11. Variable P_2_S: Nota Promedio de Materias Aprobadas en el Segundo Semestre

La variable cuantitativa real P_2_S representa la nota promedio de los estudiantes en el segundo semestre con crédito de IECE. Esta variable puede tomar valores de 6 a 10.

2.3.12. Variable P_3_S: Nota Promedio de Materias Aprobadas en el Tercer Semestre

La variable cuantitativa real P_3_S representa la nota promedio de los estudiantes en el tercer semestre con crédito de IECE. Esta variable puede tomar valores de 6 a 10.

2.7.13. Variable P_4_S: Nota Promedio de Materias Aprobadas en el Cuarto Semestre

La variable cuantitativa real P_4_S representa la nota promedio de los estudiantes en el cuarto semestre con crédito de IECE. Esta variable puede tomar valores de 6 a 10.