

UNIVERSIDAD NACIONAL DE
INGENIERÍA
FACULTAD DE CIENCIAS
Escuela Profesional de Matemática



Informe de prácticas pre-profesionales

“Análisis exploratorio de los datos de asistencia
a clases de reforzamiento de los estudiantes de
la Universidad Tecnológica del Perú”

Realizado en: Universidad Tecnológica del Perú

Tema: Tutoría

Por: Jael David Laiza Gomez

Código: 20204038K

27 de octubre de 2025

Índice general

Resumen	3
Introducción	5
Antecedentes	7
1. Objetivos	9
1.1. Generales	9
1.2. Específicos	9
2. Fundamento teórico	11
2.1. Modelo matemático	11
2.2. Estimaciones	13
3. Procedimiento empleado	17
3.1. Extracción	18
3.2. Transformación	19
3.3. Carga	22
4. Resultados	23
4.1. Tipo de sesión	24
4.2. Cursos dictados	25
4.3. Inscritos y participantes	27
4.3.1. Alumnos Inscritos	27
4.3.2. Alumnos Participantes	29
4.3.3. Inscritos-Participantes	31
4.4. Duración de sesiones	33
4.5. Puntualidad	35
4.6. Extra: Correlación entre variables numéricas	37

5. Discusión	39
5.1. Tipo	39
5.2. Curso	40
5.3. Inscritos y participantes	41
5.3.1. Inscritos	41
5.3.2. Participantes	41
5.3.3. Inscritos-Participantes	42
5.4. Duración de las sesiones	44
5.5. Puntualidad	44
6. Conclusiones	47
7. Anexos	49
7.1. Miscelánea	49
7.2. Probabilidad	49
7.3. Métricas estadísticas	50

Resumen

El contenido principal del informe se centra en realizar un análisis exploratorio de los datos obtenidos a partir de las sesiones de talleres y tutorías dictadas a los estudiantes de la *Universidad Tecnológica del Perú*, con el fin de determinar si existen variables aleatorias que modelen el comportamiento estadístico de los datos. Se presenta el esquema matemático del modelo y los resultados experimentales obtenidos a partir de los datos usando POWER BI, POWERQUERY y PYTHON.

Introducción

El presente informe corresponde a las Prácticas Pre-Profesionales realizadas en la Universidad Tecnológica del Perú (UTP), desde el 3 de marzo de 2025 hasta el 2 de agosto de 2025.

La empresa Universidad Tecnológica del Perú se dedica a formar profesionales en áreas como ingeniería, arquitectura, ciencias de la salud, ciencias sociales y ciencias de la comunicación.

En particular, la sede en cuestión a tratar (*UTP - Lima Centro*) está localizada en *Av. Arequipa 265, Lima 15046*.

El trabajo de practicante pre-profesional corresponde al de *Tutor* en la Universidad Tecnológica del Perú y tiene como función el dictado de clases de reforzamiento en cursos de matemática a los estudiantes de la misma. Esto beneficia a los estudiantes de la sede *UTP - Lima Centro* (presencial) y a estudiantes de las diferentes sedes del país (virtual).

La permanencia en el cargo de Tutor es determinada por la institución según el cumplimiento del reglamento y la calificación de satisfacción que los estudiantes asignan al Tutor de turno en las sesiones mediante encuestas.

Antecedentes

En la Universidad Tecnológica del Perú, en vista de la necesidad de brindar apoyo académico a los estudiantes en cursos de matemática o física se realizan sesiones extracurriculares para todos los estudiantes que se clasifican en *talleres* y *tutorías*. Las sesiones son no obligatorias, por lo que los estudiantes pueden optar por inscribirse o no a este tipo de actividad y, estando inscritos, pueden o no optar por asistir. Estas sesiones pueden ser dictadas en forma *presencial* o *virtual* (mediante la plataforma *Zoom*).

El objetivo general en ambos tipos de sesiones es la de apoyar a los estudiantes en la resolución de problemas relacionados a los temas vistos en las clases de los correspondientes cursos. Sin embargo, la diferencia entre ambos tipos de sesiones radica en el tiempo y el alcance.

Por ejemplo, los *talleres* presentan una duración de 90 min y un alcance máximo de hasta 100 alumnos, mientras que las *tutorías* presentan una duración de 45 min y un alcance máximo de hasta 5 alumnos. Usualmente esto permite que un *taller* se enfoque principalmente en el desarrollo de la solución de problemas mientras que una *tutoría* tiene un enfoque más personalizado para el alumno.

Hasta este punto, si bien las variables más resaltantes podrían ser las del número de alumnos inscritos y el número de alumnos asistentes, si deseamos realizar un análisis de demanda por curso, fecha y hora podría no ser suficiente. De hecho, estas últimas tendrían que ser variables que formen parte del análisis.

Dado que se encuentran datos más detallados para las sesiones dictadas en forma *virtual*, tales como datos de todas las variables mencionadas en el párrafo anterior, se considerará esto como fuente principal para el análisis exploratorio de los datos. Asimismo, se presenta la construcción matemática del problema a estudiar mediante el uso de notaciones y resultados conocidos.

Capítulo 1

Objetivos

1.1. Generales

- Realizar un análisis exploratorio de los datos de asistencia y rendimiento de cursos dictados.
- Determinar si existen variables aleatorias que modelen el comportamiento estadístico de los datos.
- Estudiar mediante modelos matemáticos la relación entre variables.

1.2. Específicos

- Realizar un proceso ETL simple para analizar los datos de asistencia de alumnos a sesiones.
- Mostrar gráficos de barras en relación a los cursos y tipos de sesiones dictadas.
- Determinar la distribución del número de alumnos inscritos a sesiones.
- Determinar la distribución del número de alumnos asistentes a sesiones.
- Determinar la relación entre el número de alumnos inscritos y el número de alumnos asistentes a sesiones.
- Determinar la distribución del número de alumnos asistentes dado un número dado el número de alumnos asistentes.
- Determinar la distribución de otras cantidades involucradas.

Capítulo 2

Fundamento teórico

2.1. Modelo matemático

Antes de estudiar las variables del problema, las planteamos matemáticamente. Para ello presentamos notaciones adecuadas y convenientes para este contexto.

Notación 2.1.1. Denotamos al conjunto de enteros en el intervalo $I \subset \mathbb{R}$ como $I_{\mathbb{Z}} := I \cap \mathbb{Z}$.

Definición 2.1.2 (Fecha y hora). Denotamos al conjunto de todas las fechas representables (en un ordenador) como

$$\mathcal{D} := \{(y, m, d) : y \in [1900; 2099]_{\mathbb{Z}}, m \in [1; 12]_{\mathbb{Z}}, d \in D(y, m)\}$$

donde $D(y, m) \subset [1, 31]_{\mathbb{Z}}$ es el conjunto de días en el mes m y año y . Además, definimos el conjunto de horas del día (con precisión de minutos) como

$$\mathcal{H} := 24\mathbb{Z} \times 60\mathbb{Z}$$

Definición 2.1.3 (Cursos). Definimos el conjunto de cursos como

$$\mathcal{C} := \{c_i : i = 1, \dots, n\}$$

para algún $n \in \mathbb{N}$, donde cada c_i representa el nombre textual (o algún tipo de identificador numérico) de un curso.

Definición 2.1.4 (Tipo de sesión). Definimos el conjunto de los tipos de sesiones como

$$\mathcal{T}_{\mathcal{S}} := \{\text{Taller, Tutoría}\}$$

cuyos elementos representan el nombre textual de los tipos de sesiones.

Definición 2.1.5 (Sesión). Definimos el conjunto de sesiones como

$$\mathcal{S} := \{s \mid s = (t_s, c_s, d_s, h_s), t_s \in \mathcal{T}_S, c_s \in \mathcal{C}, d_s \in \mathcal{D}_S, h_s \in \mathcal{H}_S\}$$

donde $\mathcal{D}_S \subset \mathcal{D}$ y $\mathcal{H}_S \subset \mathcal{H}$.

Definición 2.1.6 (Número experimental de alumnos inscritos y asistentes). Sea $s \in \mathcal{S}$ una sesión, definimos el par inscritos-asistentes de la sesión s como $\#s := (I_s, A_s) \in \mathbb{N}_0^2$ siendo I_s el número (experimental) de inscritos y A_s el número (experimental) de asistentes.

Definición 2.1.7 (Número aleatorio de alumnos inscritos y asistentes). Definimos el número de alumnos inscritos como la variable aleatoria $\mathcal{I} : \Omega \rightarrow \mathbb{N}_0$ y el número de alumnos asistentes como la variable aleatoria $\mathcal{A} : \Omega \rightarrow \mathbb{N}_0$.

En principio, \mathcal{I} y \mathcal{A} son variables aleatorias con distribuciones desconocidas. Sin embargo, dado el contexto del problema, podemos aproximar sus distribuciones a partir de los datos obtenidos.

La aproximación es posible gracias a la ley fuerte de los grandes números (Teorema 7.2.8), pues resulta de una aplicación particular a una variable aleatoria.

Ejemplo 2.1.8 (Aproximación de la distribución de probabilidad). Sea $Y : \Omega \rightarrow \mathbb{R}$ una variable aleatoria y sea X_1, X_2, X_3, \dots una sucesión de variables aleatorias independientes e idénticamente distribuidas con $X_i \sim Y$. Dado $A \subset \mathbb{R}$ fijo, determinamos la distribución de $\mathbb{1}_A(X_i)$ como

$$\mathbb{P}(\mathbb{1}_A(X_i) = x) = \begin{cases} \mathbb{P}(Y \in A) & , x = 1 \\ \mathbb{P}(Y \in A^c) & , x = 0 \end{cases}$$

De aquí, es claro que $\mathbb{E}[\mathbb{1}_A(X_i)] = \mathbb{P}(Y \in A) < \infty$ y si definimos

$$\bar{Z}_n = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_A(X_i), \quad n \in \mathbb{N}$$

entonces por la ley fuerte de los grandes números se obtiene

$$\mathbb{P}\left(\lim_{n \rightarrow \infty} \bar{Z}_n = \mathbb{E}[\mathbb{1}_A(X_i)]\right) = 1 \implies \mathbb{P}\left(\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \mathbb{1}_A(X_i) = \mathbb{P}(Y \in A)\right) = 1$$

Esto último permite fundamentar el por qué funciona el método de Montecarlo para aproximar la distribución de una variable aleatoria. En la práctica, esto es típicamente aplicado mediante histogramas sobre un conjunto de datos.

Lo que nos dice es que si deseamos aproximar $\mathbb{P}(Y \in A)$ podemos realizar n experimentos tal que en cada uno de estos registremos el valor obtenido de la variable aleatoria X_i , luego procederíamos a contabilizar cuántas veces sucede $X_i \in A$ y finalmente nuestra aproximación sería

$$\mathbb{P}(Y \in A) \approx \frac{1}{n} \sum_{i=1}^n \mathbb{1}_A(X_i) \quad (2.1)$$

para un n suficientemente grande. Es decir que a más experimentos mejor será la aproximación.



Ahora, con 2.1 estamos en condiciones de poder realizar las aproximaciones necesarias para el análisis exploratorio de los datos.

Observación 2.1.9. *Aproximamos las distribuciones de \mathcal{I} y \mathcal{A} mediante*

$$\mathbb{P}(\mathcal{I} = n) \approx \frac{\sum_{s \in \mathcal{S}} \mathbb{1}_{\{n\}}(I_s)}{n(\mathcal{S})}, \quad \mathbb{P}(\mathcal{A} = n) \approx \frac{\sum_{s \in \mathcal{S}} \mathbb{1}_{\{n\}}(A_s)}{n(\mathcal{S})} \quad (2.2)$$

para $n(\mathcal{S})$ suficientemente grande. Esto representa un método de aproximación mediante Montecarlo.

Como $n(\mathcal{S})$ representa la cantidad de sesiones (o el tamaño de \mathcal{S}), al referirnos a que este sea *suficientemente grande* damos por entendido que a mayor cantidad de sesiones existentes mayor será la calidad de la aproximación.

Definición 2.1.10 (Otras variables aleatorias). Definimos las variables aleatorias

- $T_{\mathcal{S}} : \Omega \longrightarrow \mathcal{T}_{\mathcal{S}}$, tipo de sesión;
- $C : \Omega \longrightarrow \mathcal{C}$, curso;
- $D_{\mathcal{S}} : \Omega \longrightarrow \mathcal{D}$, fecha;
- $H_{\mathcal{S}} : \Omega \longrightarrow \mathcal{H}$, hora.

Debemos hacer la distinción entre estas variables aleatorias y sus homónimos definidos anteriormente, ya que estos últimos son objetos fijos. Nuevamente, asumimos que podemos aproximar sus distribuciones mediante una aproximación de Montecarlo como en 2.2.

Notación 2.1.11. Denotamos $\hat{\mathbb{P}}_A(k) := \frac{\sum_{s \in \mathcal{S}} \mathbb{1}_{\{k\}}(a_s)}{n(\mathcal{S})}$ para $k \in A$ y $a_s \in A$ una de las componentes de $s \in \mathcal{S}$.

Observación 2.1.12. *Aproximamos las distribuciones de $T_{\mathcal{S}}$, C , $D_{\mathcal{S}}$ y $H_{\mathcal{S}}$ mediante*

$$\mathbb{P}(T_{\mathcal{S}} = t) \approx \hat{\mathbb{P}}_{\mathcal{T}_{\mathcal{S}}}(t), \quad \mathbb{P}(C = c) \approx \hat{\mathbb{P}}_{\mathcal{C}}(c), \quad \mathbb{P}(D_{\mathcal{S}} = d) \approx \hat{\mathbb{P}}_{\mathcal{D}_{\mathcal{S}}}(d), \quad \mathbb{P}(H_{\mathcal{S}} = h) \approx \hat{\mathbb{P}}_{\mathcal{H}_{\mathcal{S}}}(h) \quad (2.3)$$

para $n(\mathcal{S})$ suficientemente grande. Esto representa un método de aproximación mediante Montecarlo.

2.2. Estimaciones

Hasta este punto tenemos modeladas las variables en forma matemática, sin embargo, es requerido tener datos experimentales para obtener conclusiones. De hecho, el conjunto \mathcal{S} representa el

conjunto de datos obtenidos en este contexto, por lo que tener un \mathcal{S} suficientemente grande permitiría obtener mejores aproximaciones de las distribuciones de probabilidad.

Más exactamente, \mathcal{S} representa el conjunto de datos de las sesiones dictadas en modalidad virtual y para obtener estos datos es necesario realizar su respectiva extracción. (La etapa de extracción se detallará posteriormente).

Observación 2.2.1. *La distribución de probabilidad de la variable aleatorias número de inscritos en un curso dado es simplemente la probabilidad condicionada de \mathcal{I} dado C . Por ejemplo, si deseamos calcular la probabilidad de que hayan n inscritos en el curso c calculamos*

$$\mathbb{P}_{C=c}(\mathcal{I} = n) := \mathbb{P}(\mathcal{I} = n \mid C = c)$$

Notación 2.2.2. En motivación de la Observación 2.2.1 usamos la notación

$$\mathbb{P}_B(A) := \mathbb{P}(A \mid B) \quad (2.4)$$

para A y B eventos de un espacio de probabilidad.

Observación 2.2.3. *Dado que estamos interesados en analizar el comportamiento de la variable aleatoria \mathcal{I} dado C podríamos aprovechar esto para realizar una menor cantidad de cálculos en la aproximación de la distribución de \mathcal{I} . Es decir, por la ley de la probabilidad total es cierto que*

$$\mathbb{P}(\mathcal{I} = n) = \sum_{c \in \mathcal{C}} \mathbb{P}(C = c) \mathbb{P}(\mathcal{I} = n \mid C = c) \quad (2.5)$$

y por lo tanto podemos obtener la aproximación de la distribución de \mathcal{I} conociendo las aproximaciones de las distribuciones de C e \mathcal{I} dado C . Esto es porque para aproximar $\mathbb{P}(C = c)$ y $\mathbb{P}(\mathcal{I} = n \mid C = c)$ debemos realizar un conteo para cada $c \in \mathcal{C}$ sobre el conjunto de datos y si realizáramos el cálculo de $\mathbb{P}(\mathcal{I} = n)$ por separado tendríamos que volver a contabilizar; sin embargo, usando 2.5 podemos determinarla directamente sumando y multiplicando.

Como trabajamos bajo el supuesto de que ciertas variables del problema se comportan como variables aleatorias, entonces también deberíamos ser capaces de estimar su media y su varianza, por ejemplo. La estimación de ambas puede ser realizada mediante la maximización de la verosimilitud, pero para ello sería necesario conocer la distribución (teórica) de las mismas.

Nuevamente, mediante una aproximación de Montecarlo será posible. De hecho, esto resulta en la fórmula más usual para calcular la media y varianza de un conjunto de datos.

Dado que el caso de la estimación de la media está dada en el mismo enunciado del teorema, se muestra un ejemplo en relación a la estimación de la varianza.



Ejemplo 2.2.4 (Aproximación de la varianza). Sea $Y : \Omega \rightarrow \mathbb{R}$ una variable aleatoria y sea X_1, X_2, X_3, \dots una sucesión de variables aleatorias independientes e idénticamente distribuidas con $X_i \sim Y$ con media μ y varianza σ^2 . Sea $Z_i = (X_i - \mu)^2 \sim Z = (X - \mu)^2$, entonces

$$\mathbb{E}[Z_i] = \mathbb{E}[Z] = \mathbb{E}[(X - \mu)^2] = \text{Var}(X) = \sigma^2 < \infty$$

y si definimos

$$\bar{Z}_n = \frac{1}{n} \sum_{i=1}^n Z_i, \quad n \in \mathbb{N}$$

entonces por la ley fuerte de los grandes números se obtiene

$$\mathbb{P}\left(\lim_{n \rightarrow \infty} \bar{Z}_n = \mathbb{E}[Z_i]\right) = 1 \implies \mathbb{P}\left(\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2 = \sigma^2\right) = 1$$

Entonces si deseamos aproximar σ^2 podemos realizar n experimentos tal que en cada uno de estos registremos el valor obtenido de la variable aleatoria X_i , luego procederíamos a promediar los $(X_i - \mu)^2$ y finalmente nuestra aproximación sería

$$\sigma^2 \approx \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2 \quad (2.6)$$

para un n suficientemente grande. Es decir que a más experimentos mejor será la aproximación.

Si bien podemos contentarnos con la expresión obtenida en 2.6, en el contexto del análisis de estimadores estadísticos la expresión del lado derecho resulta ser un estimador insesgado. Para evitar esto, lo usual es cambiar el factor $\frac{1}{n}$ por $\frac{1}{n-1}$.

En resumen, los estimadores de la media μ y varianza σ^2 de una variable aleatoria real X con son

- **Media aritmética:** $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n X_i$
- **Varianza:**
 - $\hat{\sigma}_p^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2$ (poblacional)
 - $\hat{\sigma}_s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \mu)^2$ (muestral)

respectivamente. Además, cuando el tamaño n de los datos es suficientemente grande la diferencia entre la media poblacional y muestral se hace ínfima, por lo que en tales casos la elección del estimador es indistinta.

Capítulo 3

Procedimiento empleado

Existen varios procedimientos para la recopilación de datos y tienen como objetivo obtener un conjunto de datos que refleje la situación de interés (limpio) y que pueda ser utilizado para el análisis.

En particular, se muestra el procedimiento de recopilación de datos las sesiones dictadas en la *Universidad Tecnológica del Perú* mediante *ETL* (Extracción-Transformación-Carga).

3.1. Extracción

Para concretar el análisis es necesario conocer los elementos de \mathcal{S} (sesiones), es decir, contar con el conjunto de datos de interés. Por ello es necesario realizar la extracción de los datos.

Desde la plataforma [UTP] es posible acceder al historial de todas las sesiones asignadas, detallando el curso, el tipo, la fecha y hora y el número de alumnos inscritos de cada sesión (si se cuenta con acceso). Sin embargo, no es posible acceder al número de alumnos asistentes, por lo que se busca otra forma de extraer los datos. De hecho, incluso si estos últimos datos fueran accesibles desde esta plataforma aún se tendría un problema: no hay forma de exportar los datos.

Estos inconvenientes pueden ser fácilmente resueltos accediendo al historial de reuniones de *Zoom*, ya que las sesiones virtuales se realizan en esta plataforma. Aquí es posible extraer los datos en múltiples archivos de formato .csv donde cada archivo contiene los datos de las sesiones en relación a un mes específico y además cierto tipo de información de la sesión.

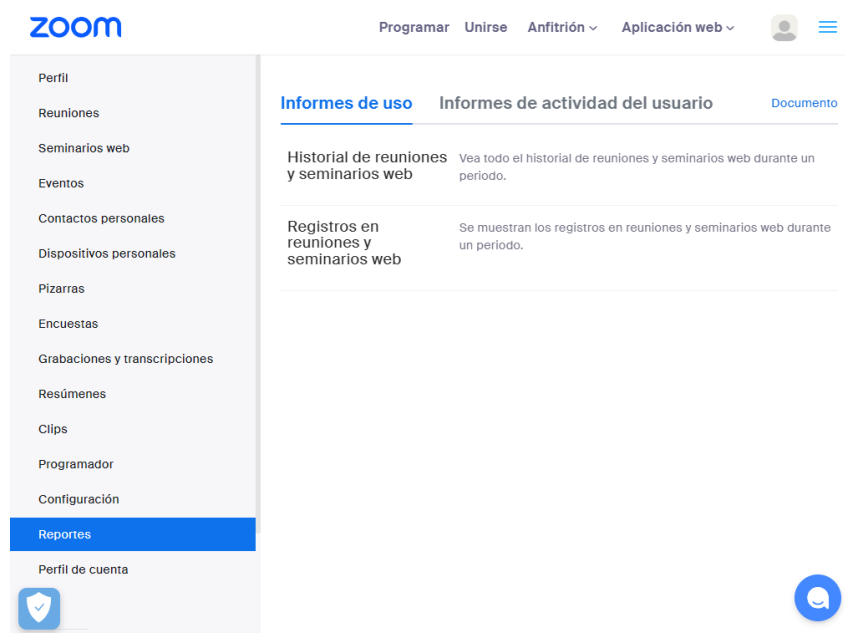


Figura 3.1: Fuente de extracción de los datos (apartado de reportes en *Zoom*) [Zoo].

Específicamente, para cada mes se tienen dos archivos .csv, donde cada uno contiene su respectiva tabla:

1. *registrationMeetings*, que contiene los datos generales de la sesiones programadas (datos previos);
2. *usermeeting*, los datos de actividad de las sesiones programadas (datos posteriores y algunos



3.2. TRANSFORMACIÓN

datos redundantes como columnas de *registrationMeetings*, entre otros);

Se aclara que en los datos de una sesión programada es imposible conocer el número de asistentes, sin embargo en los datos de actividad de las sesiones programadas sí (pues ya sucedió un número de asistentes).

Entonces, si bien no contamos directamente con un único archivo para agrupar la totalidad los datos, mediante un proceso de transformación de los mismos es posible hacerlo. Por ello nos es suficiente contar con los archivos .csv extraídos.








	 AC ID	 AC Tipo	 AC Curso	 Fecha	 Hora	 123 Inscripciones
	<div><div></div> Válido 100 %</div> <div><div></div> Error 0 %</div> <div><div></div> Vacío 0 %</div>	<div><div></div> Válido 100 %</div> <div><div></div> Error 0 %</div> <div><div></div> Vacío 0 %</div>	<div><div></div> Válido 100 %</div> <div><div></div> Error 0 %</div> <div><div></div> Vacío 0 %</div>	<div><div></div> Válido 100 %</div> <div><div></div> Error 0 %</div> <div><div></div> Vacío 0 %</div>	<div><div></div> Válido 100 %</div> <div><div></div> Error 0 %</div> <div><div></div> Vacío 0 %</div>	<div><div></div> Válido 100 %</div> <div><div></div> Error 0 %</div> <div><div></div> Vacío 0 %</div>
1	885 1636 2512	Taller	GEOMETRIA	4/06/2025	1:00:00 p. m.	1
2	874 1998 5448	Taller	GEOMETRIA	4/06/2025	2:45:00 p. m.	2
3	839 4685 8633	Taller	GEOMETRIA	4/06/2025	5:30:00 p. m.	6
4	899 6509 0022	Taller	GEOMETRIA	6/06/2025	1:00:00 p. m.	8
5	842 9161 9284	Taller	CÁLCULO I	6/06/2025	2:45:00 p. m.	10
6	849 6575 1141	Taller	ESTADISTICA DESCRIPTIVA Y PROB	6/06/2025	5:30:00 p. m.	36
7	873 2696 2239	Taller	CÁLCULO I	9/06/2025	1:45:00 p. m.	22
8	865 9098 2036	Taller	CÁLCULO I	9/06/2025	3:45:00 p. m.	25
9	854 7542 7796	Taller	CÁLCULO I	9/06/2025	5:30:00 p. m.	29
10	857 0704 1869	Tutoría	ESTADISTICA DESCRIPTIVA Y PROB	11/06/2025	1:15:00 p. m.	6
11	874 6746 4315	Taller	CÁLCULO I	11/06/2025	2:15:00 p. m.	56
12	868 6395 2186	Taller	CÁLCULO I	11/06/2025	4:00:00 p. m.	36
13	810 3210 1560	Tutoría	ESTADISTICA DESCRIPTIVA Y PROB	11/06/2025	5:45:00 p. m.	8
14	878 9850 6606	Taller	ESTADISTICA DESCRIPTIVA Y PROB	13/06/2025	1:45:00 p. m.	10
15	868 4748 2369	Taller	GEOMETRIA	13/06/2025	3:45:00 p. m.	4

Figura 3.2: Primeras filas de *registrationMeetings*

A ^B _C ID	A ^B _C Tipo	A ^B _C Curso	Fecha	Hora Inicio	Hora Fin	i ² ₃ Participantes	i ² ₃ Duración (minutos)	i ² ₃ Total de minutos de los participantes
● Válido 100 % ● Error 0 % ● Vacío 0 %	● Válido 100 % ● Error 0 % ● Vacío 0 %	● Válido 100 % ● Error 0 % ● Vacío 0 %	● Válido 100 % ● Error 0 % ● Vacío 0 %	● Válido 100 % ● Error 0 % ● Vacío 0 %	● Válido 100 % ● Error 0 % ● Vacío 0 %	● Válido 100 % ● Error 0 % ● Vacío 0 %	● Válido 100 % ● Error 0 % ● Vacío 0 %	● Válido 100 % ● Error 0 % ● Vacío 0 %
1	839 4685 8633	Taller	GEOMETRIA	3/06/2025	9:10:16 p. m.	9:10:25 p. m.	1	1
2	874 1998 5448	Taller	GEOMETRIA	4/06/2025	2:40:30 p. m.	4:16:49 p. m.	2	97
3	885 1636 2512	Taller	GEOMETRIA	4/06/2025	12:53:47 p. m.	2:07:45 p. m.	2	74
4	818 0087 8359	Taller	GEOMETRIA	4/06/2025	2:10:45 p. m.	2:11:56 p. m.	1	2
5	839 4685 8633	Taller	GEOMETRIA	4/06/2025	5:25:15 p. m.	6:40:48 p. m.	8	76
6	874 1998 5448	Taller	GEOMETRIA	4/06/2025	4:17:39 p. m.	4:51:40 p. m.	1	35
7	849 6575 1141	Taller	ESTADISTICA DESCRIPTIVA Y ...	5/06/2025	6:21:16 p. m.	6:29:02 p. m.	1	8
8	899 6509 0022	Taller	GEOMETRIA	6/06/2025	12:55:49 p. m.	2:39:51 p. m.	12	105
9	899 6509 0022	Taller	GEOMETRIA	6/06/2025	2:40:17 p. m.	2:40:24 p. m.	1	1
10	842 9161 9284	Taller	CÁLCULO I	6/06/2025	11:43:20 a. m.	11:43:32 a. m.	1	1
11	849 6575 1141	Taller	ESTADISTICA DESCRIPTIVA Y ...	6/06/2025	5:21:55 p. m.	6:51:54 p. m.	16	90
12	849 6575 1141	Taller	ESTADISTICA DESCRIPTIVA Y ...	6/06/2025	7:00:03 p. m.	7:00:08 p. m.	1	1
13	842 9161 9284	Taller	CÁLCULO I	6/06/2025	2:40:08 p. m.	5:21:07 p. m.	7	161
14	854 7542 7796	Taller	CÁLCULO I	9/06/2025	5:21:41 p. m.	7:06:49 p. m.	23	106
15	865 9098 2036	Taller	CÁLCULO I	9/06/2025	3:39:24 p. m.	3:54:55 p. m.	5	41

Figura 3.3: Primeras filas de *usermeeting*

3.2. Transformación

Hasta este punto se cuenta con dos tipos de tablas: *registrationMeetings* y *usermeeting*. Sin embargo, necesitamos agrupar los datos previos y posteriores para cada sesión existente. Para ello

podemos recurrir a cualquier *software* (POWER BI, POWERQUERY) o lenguaje de programación (PYTHON) que incorpore la operación **LEFT JOIN**.

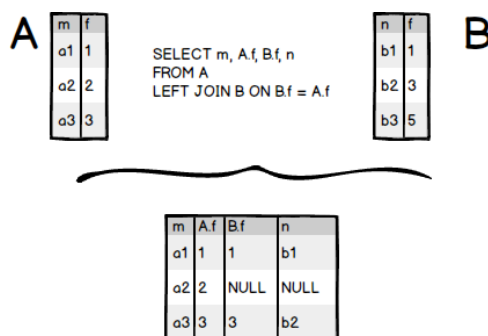


Figura 3.4: Interpretación gráfica de la operación **LEFT JOIN** entre dos tablas

Esto será posible porque *registrationMeetings* y *usermeeting* cuentan con una columna que contiene el identificador único para cada sesión (*ID* de Zoom). Llamaremos a la tabla resultante como *Meetings*.

A _C ID	A _C Tipo	A _C Curso	Fecha	Hora	Inscripciones	usermeeting_2025_0...
<ul style="list-style-type: none"> Válido 100 % Error 0 % Vacio 0 % 	<ul style="list-style-type: none"> Válido 100 % Error 0 % Vacio 0 % 	<ul style="list-style-type: none"> Válido 100 % Error 0 % Vacio 0 % 	<ul style="list-style-type: none"> Válido 100 % Error 0 % Vacio 0 % 	<ul style="list-style-type: none"> Válido 100 % Error 0 % Vacio 0 % 	<ul style="list-style-type: none"> Válido 100 % Error 0 % Vacio 0 % 	<ul style="list-style-type: none"> Válido 100 % Error 0 % Vacio 0 %
1 885 1636 2512	Taller	GEOMETRIA	4/06/2025	1:00:00 p. m.	1	Table
2 874 1998 5448	Taller	GEOMETRIA	4/06/2025	2:45:00 p. m.	2	Table
3 839 4685 8633	Taller	GEOMETRIA	4/06/2025	5:30:00 p. m.	6	Table
4 899 6509 0022	Taller	GEOMETRIA	6/06/2025	1:00:00 p. m.	8	Table
5 842 9161 9284	Taller	CÁLCULO I	6/06/2025	2:45:00 p. m.	10	Table
6 849 6575 1141	Taller	ESTADISTICA DESCRIPTIVA Y PROB	6/06/2025	5:30:00 p. m.	36	Table
7 873 2696 2239	Taller	CÁLCULO I	9/06/2025	1:45:00 p. m.	22	Table
8 865 9098 2036	Taller	CÁLCULO I	9/06/2025	3:45:00 p. m.	25	Table
9 854 7542 7796	Taller	CÁLCULO I	9/06/2025	5:30:00 p. m.	29	Table
10 857 0704 1869	Tutoría	ESTADISTICA DESCRIPTIVA Y PROB	11/06/2025	1:15:00 p. m.	6	Table
11 874 6746 4315	Taller	CÁLCULO I	11/06/2025	2:15:00 p. m.	56	Table
12 868 6395 2186	Taller	CÁLCULO I	11/06/2025	4:00:00 p. m.	36	Table

Figura 3.5: Primeras filas de *Meetings*

Observación 3.2.1. Para el procedimiento de transformación de las tablas involucradas se utiliza POWER BI y POWERQUERY.

Ahora, la tarea no es tan sencilla como obtener la tabla *Meetings* y realizar el análisis con la misma, sino que puede verse involucrado un proceso de limpieza (aunque simple) del mismo. Esto es debido a que *Meetings* heredará las demás columnas de *registrationMeetings* y *usermeeting* y, en general, es posible que aparezcan elementos de la tabla con valores *NULL* (lo cual no es favorable para el análisis), filas con *ID* repetido columnas repetidas. Por tanto habrá que realizar a una limpieza de datos en la tabla *Meetings*.

La mayoría de valores *NULL* aparecen porque *registrationMeetings* contiene datos de sesiones



3.2. TRANSFORMACIÓN

que fueron inicialmente programadas pero posteriormente canceladas y que al operarse mediante **LEFT JOIN** con *usermeeting* generó datos de actividad inexistentes.

La tabla *Meetings* contiene *ID* repetidos porque en *usermeeting* existen filas con *ID* repetido (como ya se había anticipado en 3.1). La razón del por qué aparecen estas filas es porque en realidad la tabla *usermeeting* contiene los datos de actividad de las sesiones registradas por *Zoom* con sus correspondientes *ID* e independientemente de la hora programada, es decir, que se puede registrar tanto antes como después de las sesiones programadas que aparecen en *registrationMeetings* (donde todas las filas tienen *ID* único). Por lo tanto, basta que algún estudiante ingrese y salga a una sesión fuera del rango de la hora programada para generar datos de actividad de la misma en *usermeeting*.

Entonces, el objetivo principal en la limpieza de datos de *Meetings* será:

1. Eliminar las filas con datos *NULL* (ya que las sesiones no se realizaron).
2. Eliminar las columnas repetidas.
3. Eliminar las filas con *ID* repetido.

La tabla obtenida de *Meetings* posterior a la limpieza se llamará *MeetingsClean* y esta servirá de base para realizar las transformaciones que sean necesarias.

ID	Tipo	Curso	Fecha	Hora	Hora Inicio	Hora Fin	Alumnos Inscritos	Alumnos Participantes	Duración (minutos)	Total de minutos de L...
1	Valido	100 %	Valido	100 %	Valido	100 %	Valido	100 %	Valido	100 %
2	Error	0 %	Error	0 %	Error	0 %	Error	0 %	Error	0 %
3	Vacio	0 %	Vacio	0 %	Vacio	0 %	Vacio	0 %	Vacio	0 %
4	Valido	100 %	Valido	100 %	Valido	100 %	Valido	100 %	Valido	100 %
5	Error	0 %	Error	0 %	Error	0 %	Error	0 %	Error	0 %
6	Vacio	0 %	Vacio	0 %	Vacio	0 %	Vacio	0 %	Vacio	0 %
7	Valido	100 %	Valido	100 %	Valido	100 %	Valido	100 %	Valido	100 %
8	Error	0 %	Error	0 %	Error	0 %	Error	0 %	Error	0 %
9	Vacio	0 %	Vacio	0 %	Vacio	0 %	Vacio	0 %	Vacio	0 %
10	Valido	100 %	Valido	100 %	Valido	100 %	Valido	100 %	Valido	100 %
11	Error	0 %	Error	0 %	Error	0 %	Error	0 %	Error	0 %
12	Vacio	0 %	Vacio	0 %	Vacio	0 %	Vacio	0 %	Vacio	0 %
13	Valido	100 %	Valido	100 %	Valido	100 %	Valido	100 %	Valido	100 %
14	Error	0 %	Error	0 %	Error	0 %	Error	0 %	Error	0 %
15	Vacio	0 %	Vacio	0 %	Vacio	0 %	Vacio	0 %	Vacio	0 %
16	Valido	100 %	Valido	100 %	Valido	100 %	Valido	100 %	Valido	100 %
17	Error	0 %	Error	0 %	Error	0 %	Error	0 %	Error	0 %
18	Vacio	0 %	Vacio	0 %	Vacio	0 %	Vacio	0 %	Vacio	0 %
19	Valido	100 %	Valido	100 %	Valido	100 %	Valido	100 %	Valido	100 %
20	Error	0 %	Error	0 %	Error	0 %	Error	0 %	Error	0 %
21	Vacio	0 %	Vacio	0 %	Vacio	0 %	Vacio	0 %	Vacio	0 %
22	Valido	100 %	Valido	100 %	Valido	100 %	Valido	100 %	Valido	100 %
23	Error	0 %	Error	0 %	Error	0 %	Error	0 %	Error	0 %
24	Vacio	0 %	Vacio	0 %	Vacio	0 %	Vacio	0 %	Vacio	0 %
25	Valido	100 %	Valido	100 %	Valido	100 %	Valido	100 %	Valido	100 %

Figura 3.6: Primeras filas de *MeetingsClean*

Hasta este punto, para obtener los datos (y por tanto los elementos de \mathcal{S}) necesarios para el análisis es requerido realizar algunas transformaciones sobre las columnas de *MeetingsClean* sujetas principalmente a separación de caracteres en columnas que contienen texto.

Finalmente, la tabla resultante de aplicar las transformaciones a las columnas de *MeetingsClean* se llamará *MeetingsCleanInfo*.

A ₁ ID	A ₂ Tipo	A ₃ Curso	Fecha	Hora	Hora Inicio	Hora Fin	1.2 Puntualidad (minutos)	1.3 Duración (minutos)	1.4 Alumnos Inscritos	1.5 Alumnos Participantes	1.6 Participación
1	885 6586 2512	Taller	GEOMETRIA	4/06/2025	1:00:00 p. m.	12:53:47 p. m.	2:07:45 p. m.	6.22	74	1	1
2	874 1988 5448	Taller	GEOMETRIA	4/06/2025	2:45:00 p. m.	2:40:30 p. m.	4:16:49 p. m.	4.5	97	2	1
3	839 4685 8633	Taller	GEOMETRIA	4/06/2025	5:30:00 p. m.	5:25:15 p. m.	6:40:48 p. m.	4.75	76	6	7
4	899 6509 0022	Taller	GEOMETRIA	6/06/2025	1:00:00 p. m.	12:55:49 p. m.	2:39:51 p. m.	4.18	105	8	11
5	842 5555 5384	Taller	CÁLCULO I	6/06/2025	2:45:00 p. m.	2:40:08 p. m.	5:21:57 p. m.	4.87	101	6	6
6	840 6575 1141	Taller	ESTADISTICA DESCRIPTIVA Y ...	6/06/2025	5:30:00 p. m.	5:21:55 p. m.	6:51:54 p. m.	8.08	90	36	15
7	873 2696 2239	Taller	CÁLCULO I	9/06/2025	3:45:00 p. m.	3:37:23 p. m.	5:23:35 p. m.	7.62	107	22	7
8	865 9098 2036	Taller	CÁLCULO I	9/06/2025	3:45:00 p. m.	3:39:24 p. m.	5:54:55 p. m.	5.6	16	25	4
9	854 7542 7796	Taller	CÁLCULO I	9/06/2025	5:30:00 p. m.	5:21:41 p. m.	7:06:49 p. m.	8.32	106	29	22
10	857 0704 1889	Tutoría	ESTADISTICA DESCRIPTIVA Y ...	11/06/2025	2:15:00 p. m.	1:05:32 p. m.	1:55:16 p. m.	9.47	50	6	6
11	874 6746 4115	Taller	CÁLCULO I	11/06/2025	2:15:00 p. m.	2:03:50 p. m.	3:42:25 p. m.	11.17	99	36	48
12	868 6395 2186	Taller	CÁLCULO I	11/06/2025	4:00:00 p. m.	3:44:11 p. m.	5:27:52 p. m.	15.82	104	36	23
13	810 3210 1560	Tutoría	ESTADISTICA DESCRIPTIVA Y ...	11/06/2025	5:45:00 p. m.	5:37:18 p. m.	6:13:27 p. m.	7.7	37	8	5
14	878 9850 6606	Taller	ESTADISTICA DESCRIPTIVA Y ...	13/06/2025	3:45:00 p. m.	3:35:44 p. m.	5:01:08 p. m.	9.27	88	10	8
15	868 4748 2369	Taller	GEOMETRIA	13/06/2025	3:45:00 p. m.	3:36:49 p. m.	4:48:21 p. m.	8.18	72	4	3
16	821 1137 0485	Taller	MATEMATICA I	13/06/2025	5:30:00 p. m.	5:22:05 p. m.	6:55:33 p. m.	7.92	94	55	49
17	855 5915 5115	Tutoría	MATEMATICA I	14/06/2025	9:15:00 a. m.	9:06:35 a. m.	9:58:51 a. m.	8.42	53	15	3
18	890 6385 7988	Tutoría	ESTADISTICA DESCRIPTIVA Y ...	14/06/2025	10:15:00 a. m.	10:05:16 a. m.	10:47:57 a. m.	8.79	43	5	3
19	878 1970 2961	Taller	MATEMATICA I	14/06/2025	11:15:00 a. m.	11:06:34 a. m.	12:30:31 p. m.	8.43	94	108	62
20	870 7123 4406	Taller	CÁLCULO I	16/06/2025	3:45:00 p. m.	3:38:23 p. m.	5:16:10 p. m.	6.62	98	8	7
21	864 8856 7224	Taller	CÁLCULO I	16/06/2025	3:45:00 p. m.	3:37:08 p. m.	4:47:22 p. m.	7.87	71	13	12
22	830 4927 0059	Taller	CÁLCULO I	16/06/2025	5:30:00 p. m.	5:20:32 p. m.	6:42:20 p. m.	9.47	82	22	13
23	842 4070 1289	Tutoría	ESTADISTICA DESCRIPTIVA Y ...	18/06/2025	2:15:00 p. m.	1:05:26 p. m.	1:51:27 p. m.	9.57	47	6	0
24	898 7690 5564	Tutoría	MATEMATICA I	18/06/2025	2:15:00 p. m.	2:05:26 p. m.	3:27:49 p. m.	9.57	83	11	4
25	892 8636 5807	Taller	CÁLCULO I	18/06/2025	4:00:00 p. m.	3:50:19 p. m.	4:57:52 p. m.	9.68	68	7	6
26	842 4338 3778	Tutoría	ESTADISTICA DESCRIPTIVA Y ...	18/06/2025	5:45:00 p. m.	5:36:13 p. m.	6:37:45 p. m.	8.78	62	6	2
27	838 6605 4682	Taller	ESTADISTICA DESCRIPTIVA Y ...	20/06/2025	3:45:00 p. m.	3:36:52 p. m.	5:09:17 p. m.	8.13	93	26	16
28	812 6197 6371	Taller	GEOMETRIA	20/06/2025	3:45:00 p. m.	3:35:40 p. m.	5:03:42 p. m.	9.33	89	9	6
29	847 6524 4560	Taller	MATEMATICA I	20/06/2025	5:30:00 p. m.	5:19:23 p. m.	7:12:40 p. m.	10.62	114	61	49
30	873 9097 6975	Tutoría	MATEMATICA I	21/06/2025	9:15:00 a. m.	9:05:20 a. m.	10:03:42 a. m.	9.67	59	6	2
31	898 6225 7063	Tutoría	ESTADISTICA DESCRIPTIVA Y ...	21/06/2025	10:15:00 a. m.	10:06:31 a. m.	11:04:16 a. m.	8.48	56	10	12
32	838 2496 6602	Taller	MATEMATICA I	21/06/2025	11:15:00 a. m.	11:05:40 a. m.	12:51:46 p. m.	9.33	107	67	58
33	899 9965 2698	Taller	CÁLCULO I	23/06/2025	3:45:00 p. m.	3:37:11 p. m.	5:42:26 p. m.	7.82	66	4	3
34	827 6078 0797	Taller	CÁLCULO I	23/06/2025	3:45:00 p. m.	3:35:52 p. m.	5:17:26 p. m.	9.13	102	5	4
35	884 4008 8615	Taller	CÁLCULO I	23/06/2025	5:30:00 p. m.	5:20:54 p. m.	6:54:51 p. m.	9.1	94	3	2

Figura 3.7: Primeras filas y columnas de *MeetingsCleanInfo*

3.3. Carga

Dada la poca complejidad y el poco tamaño de los datos, esta etapa consiste únicamente en la elaboración de gráficos y/o visualizaciones con POWER BI a partir de las tablas obtenidas, principalmente *MeetingsCleanInfo*.

Adelantamos que el tamaño de la tabla *MeetingsCleanInfo* es de 147 filas, y por tanto $n(\mathcal{S}) = 147$, y 12 columnas las cuales son:

- | | | |
|----------|--------------------------|---|
| 1. ID | 6. Hora Inicio | 11. Alumnos Participantes |
| 2. Tipo | 7. Hora Fin | 12. Total de minutos de los participantes |
| 3. Curso | 8. Puntualidad (minutos) | |
| 4. Fecha | 9. Duración (minutos) | |
| 5. Hora | 10. Alumnos Inscritos | |

Asimismo, se presentan los cursos dictados:

- | | | | |
|------------------|---------------|----------------------------|-----------------------------|
| 1. Matemática I | 3. Cálculo I | 5. Estadística Descriptiva | 7. Matemática para Medicina |
| 2. Matemática II | 4. Cálculo II | 6. Geometría | |

los cuales son los elementos de \mathcal{C} .

Capítulo 4

Resultados

Los resultados obtenidos derivan de un análisis exploratorio de los datos (*EDA*) se utiliza una combinación de uso de POWER BI y PYTHON, permitiendo complementar las características y ventajas que cada uno posee. Recordemos que todo el análisis se hace a través de *MeetingsCleanInfo* que contiene los datos de interés (elementos de \mathcal{S}).

De hecho, al hacer uso de herramientas que permiten obtener visualizaciones o gráficos (histogramas) a partir de los datos podemos apoyarnos de los mismos para ver qué distribuciones tienen las variables aleatorias involucradas.

Sin embargo, en esta sección solo se presentan los gráficos y tablas sin recurrir a la interpretación de resultados.

Se resalta el hecho de que todos los archivos de datos y visualizaciones se encuentran en el repositorio de *GitHub* [[One25](#)].

4.1. Tipo de sesión

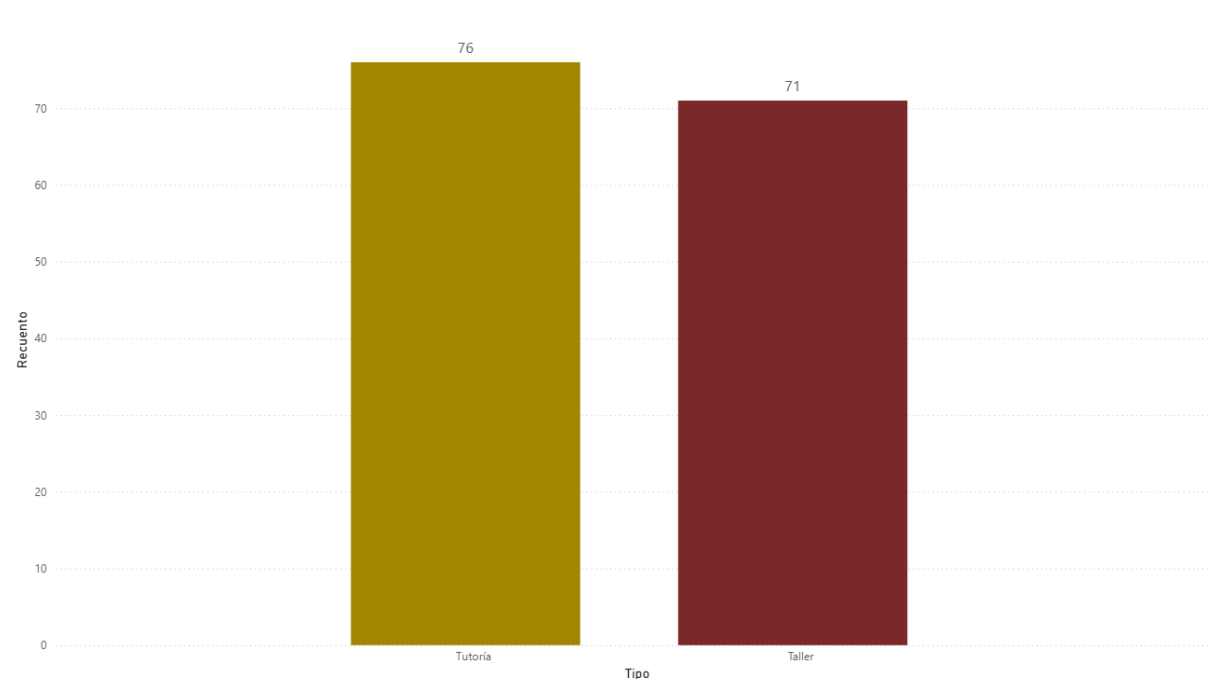


Figura 4.1: Gráfico de barras del tipo de sesión (POWER BI)

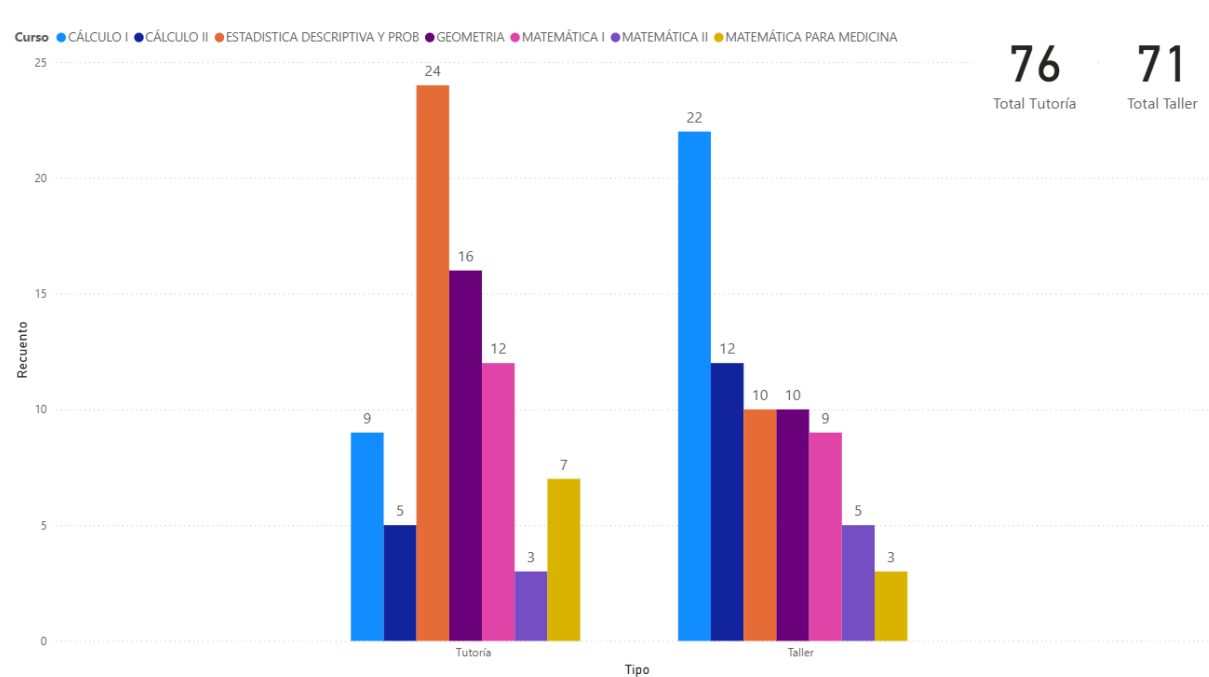


Figura 4.2: Gráfico de barras del tipo de sesión por curso (POWER BI)



4.2. Cursos dictados

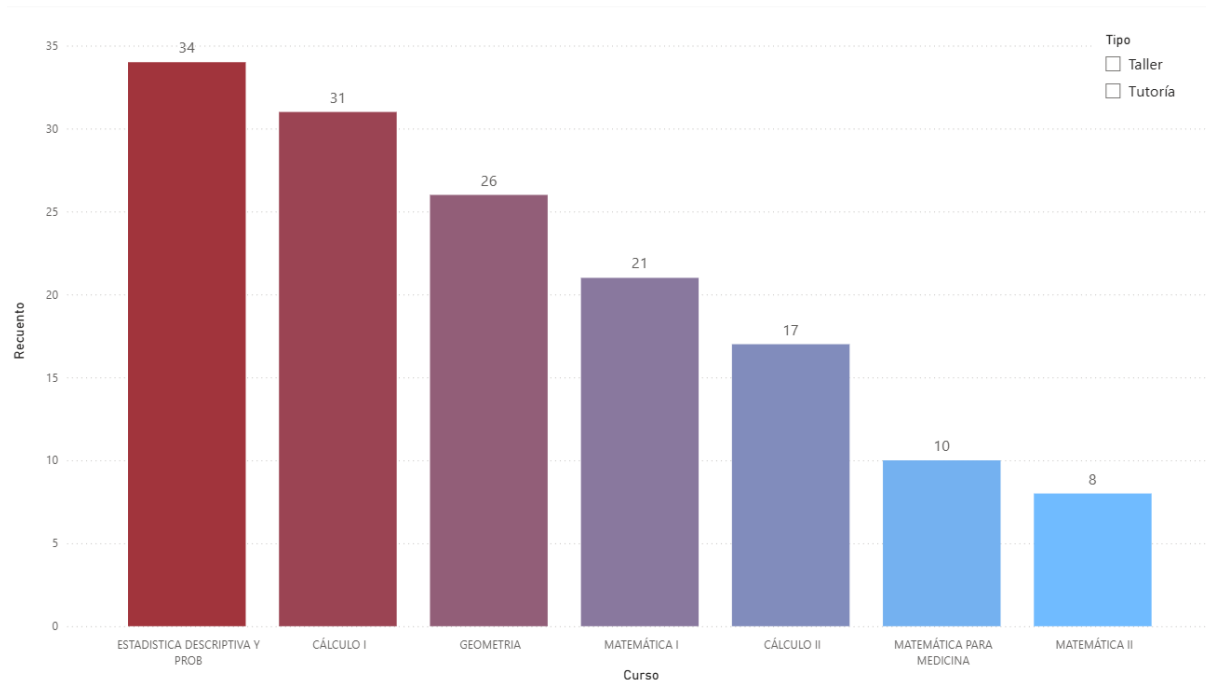


Figura 4.3: Histograma de los cursos dictados (POWER BI)

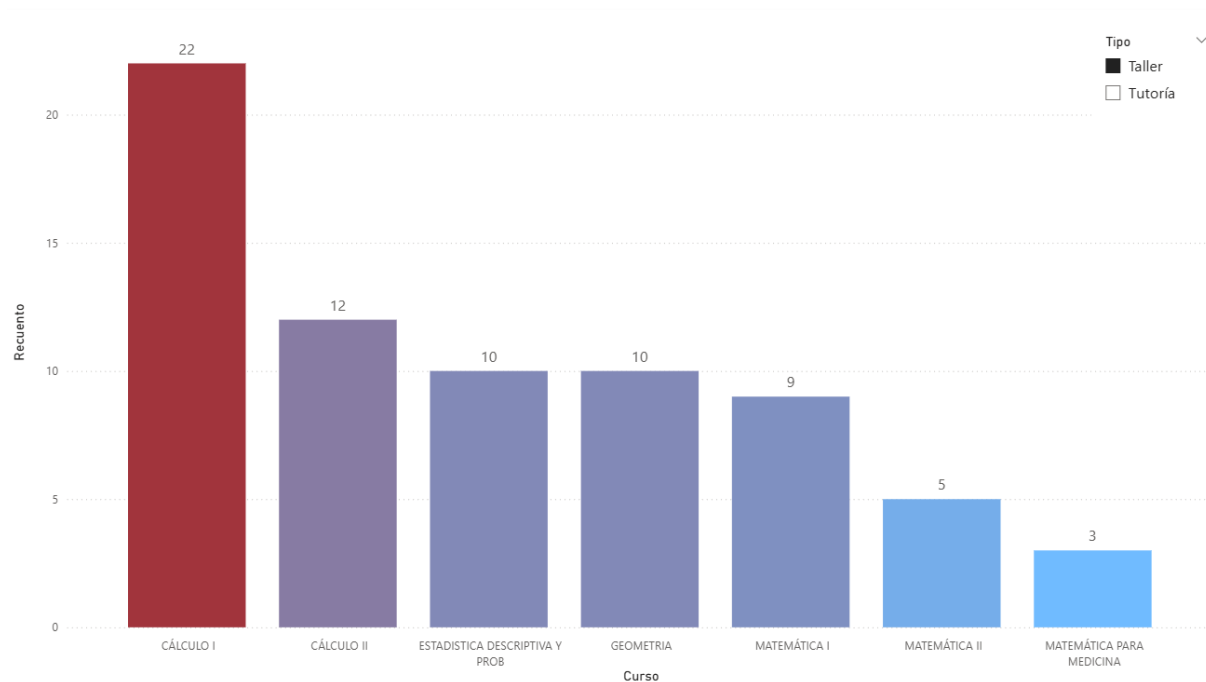


Figura 4.4: Histograma de los cursos dictados en talleres (POWER BI)

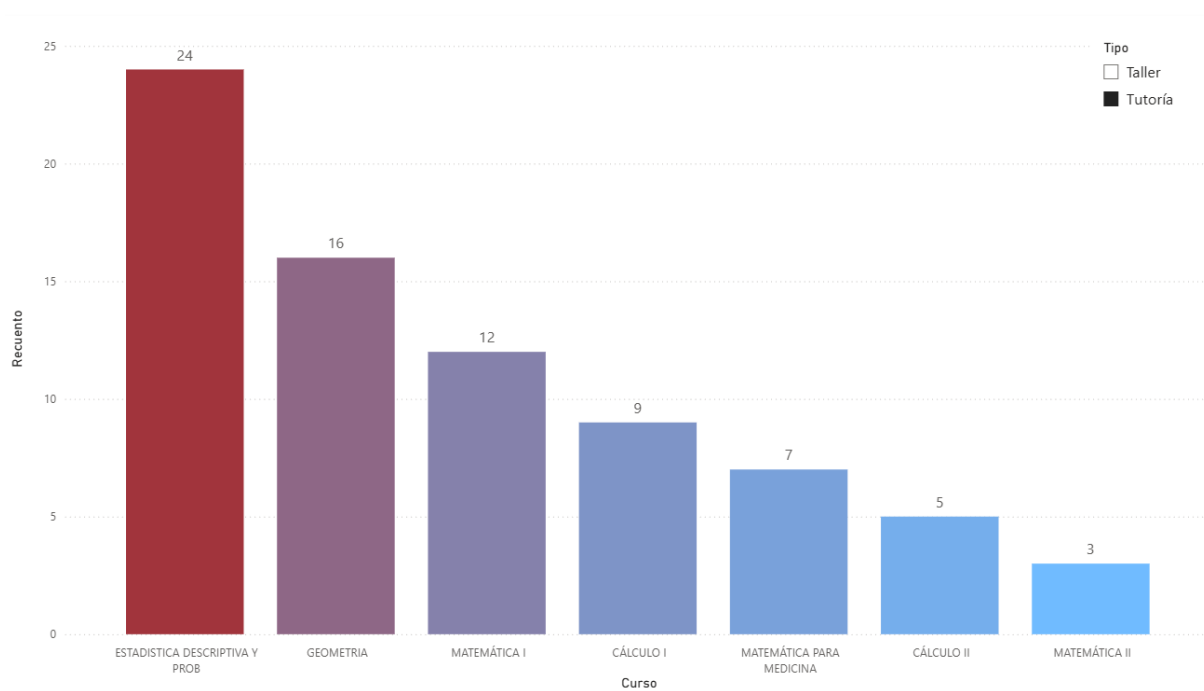


Figura 4.5: Histograma de los cursos dictados en tutorías (POWER BI)



4.3. Inscritos y participantes

4.3.1. Alumnos Inscritos

El siguiente gráfico muestra el histograma del número de alumnos inscritos en todas las sesiones, es decir el conteo sobre los valores posibles de I_s para todos los $s \in \mathcal{S}$. Además, como esta variable es de tipo numérica discreta consideramos lo siguiente al realizar la gráfica:

- No se agrupan los valores de I_s en intervalos.
- Se muestran los estadísticos de la media, desviación estándar, mediana y kurtosis.

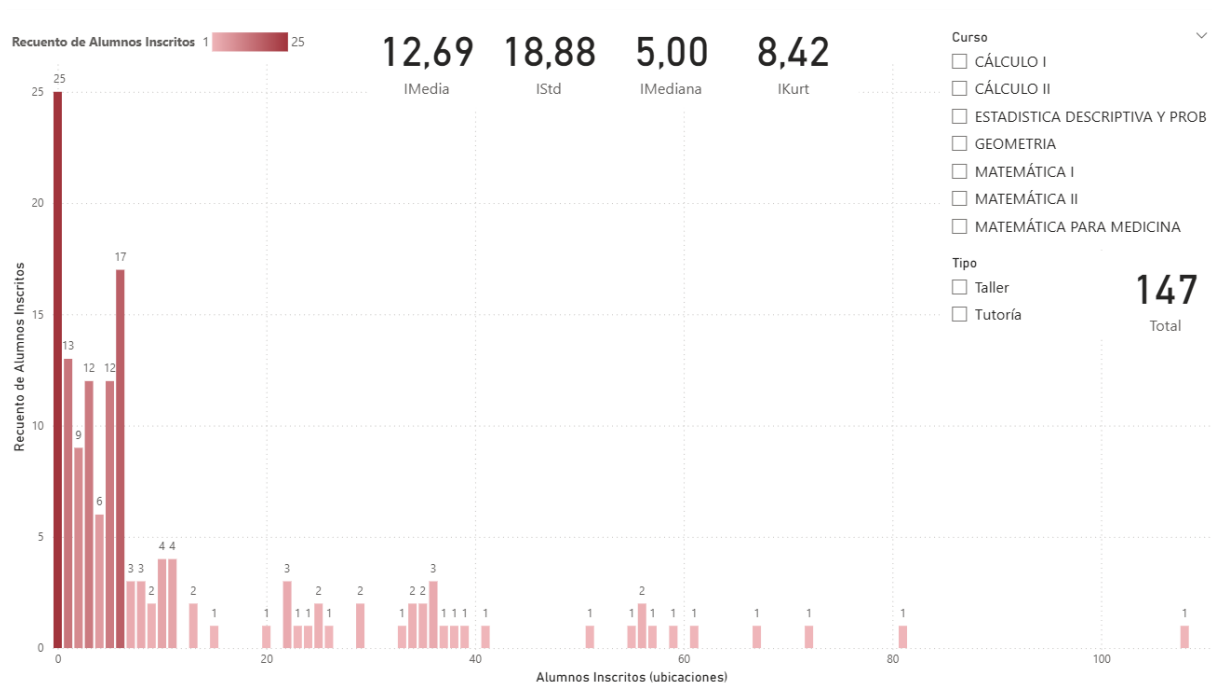


Figura 4.6: Histograma de inscritos (POWER BI)

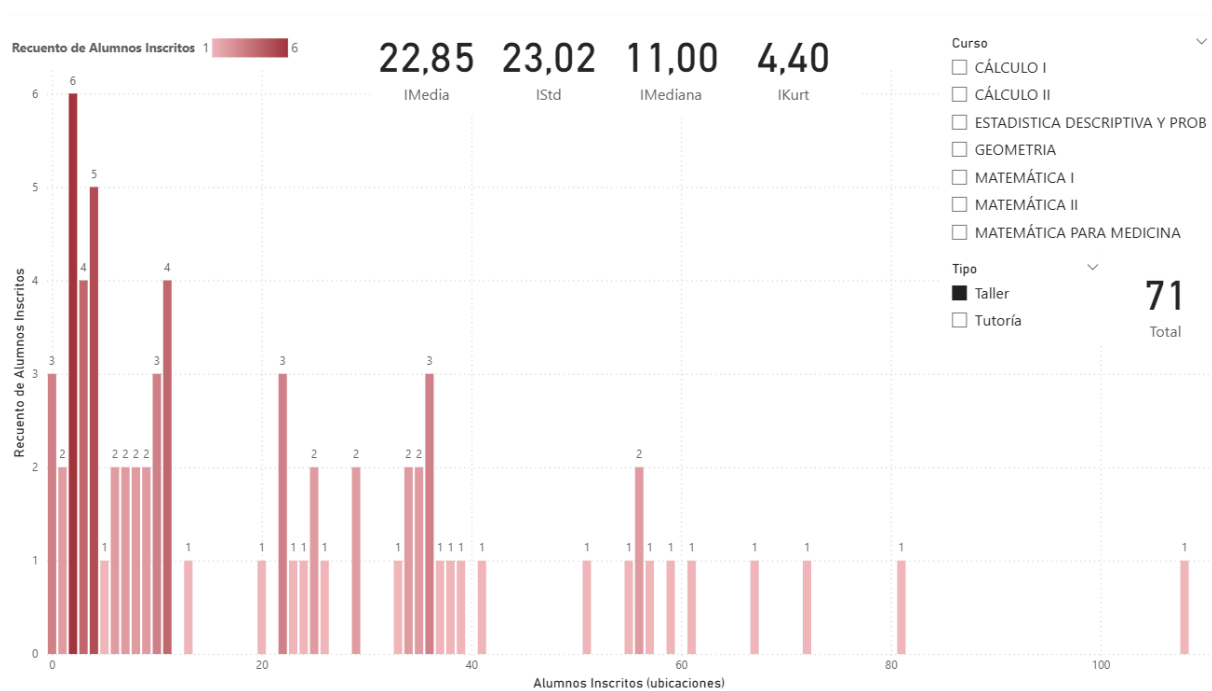


Figura 4.7: Histograma de inscritos en talleres (POWER BI)

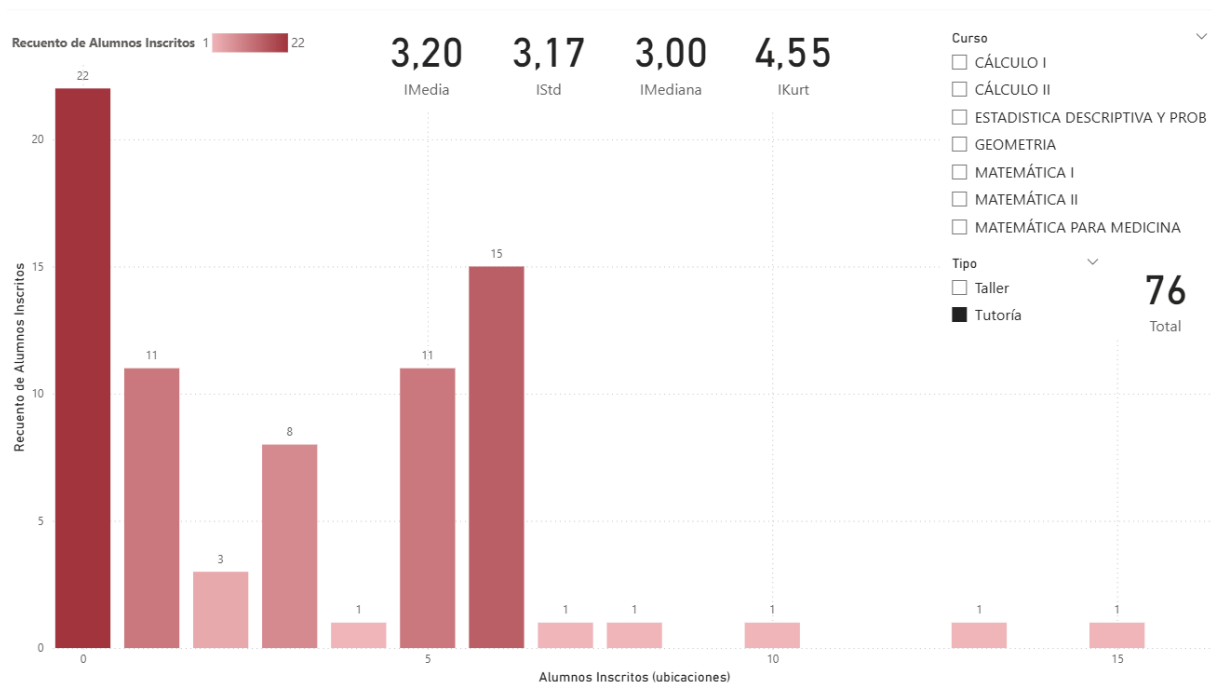


Figura 4.8: Histograma de inscritos en tutorías (POWER BI)



4.3. INSCRITOS Y PARTICIPANTES

4.3.2. Alumnos Participantes

El siguiente gráfico muestra el histograma del número de alumnos participantes en todas las sesiones, es decir el conteo sobre los valores posibles de A_s para todos los $s \in \mathcal{S}$. Además, como esta variable es de tipo numérica discreta consideramos lo siguiente al realizar la gráfica (al igual que para la Figura 4.6):

- No se agrupan los valores de I_s en intervalos.
- Se muestran los estadísticos de la media, desviación estándar, mediana y kurtosis.

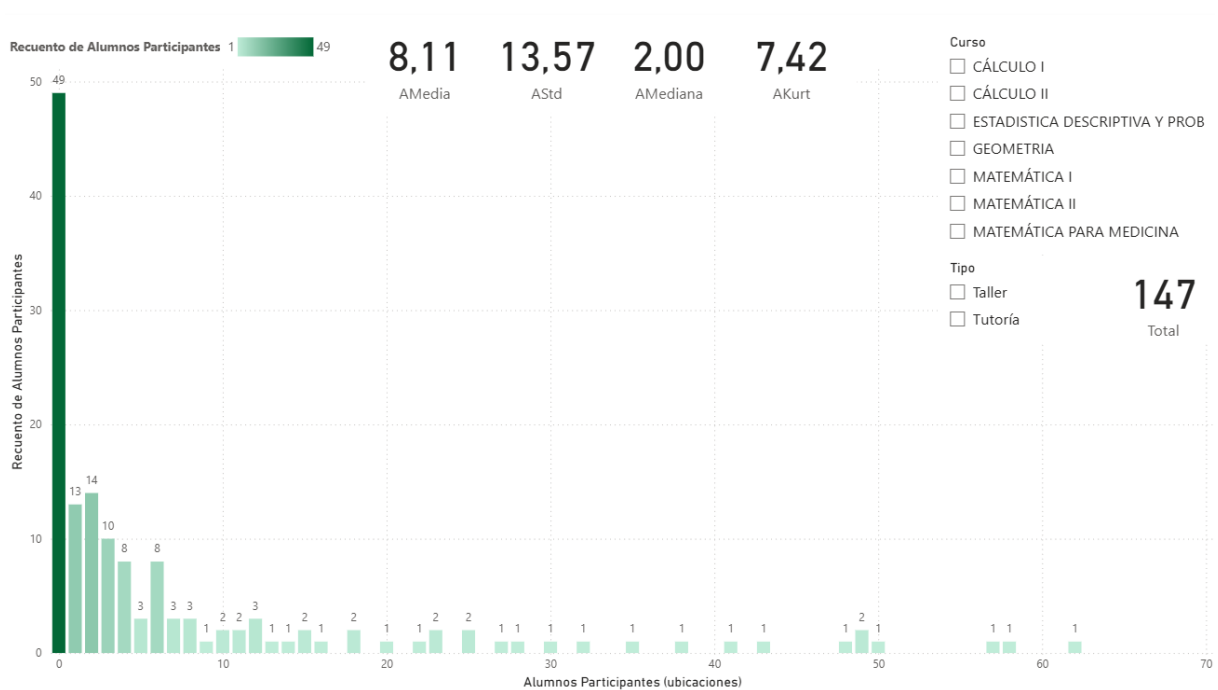


Figura 4.9: Histograma de participantes (POWER BI)

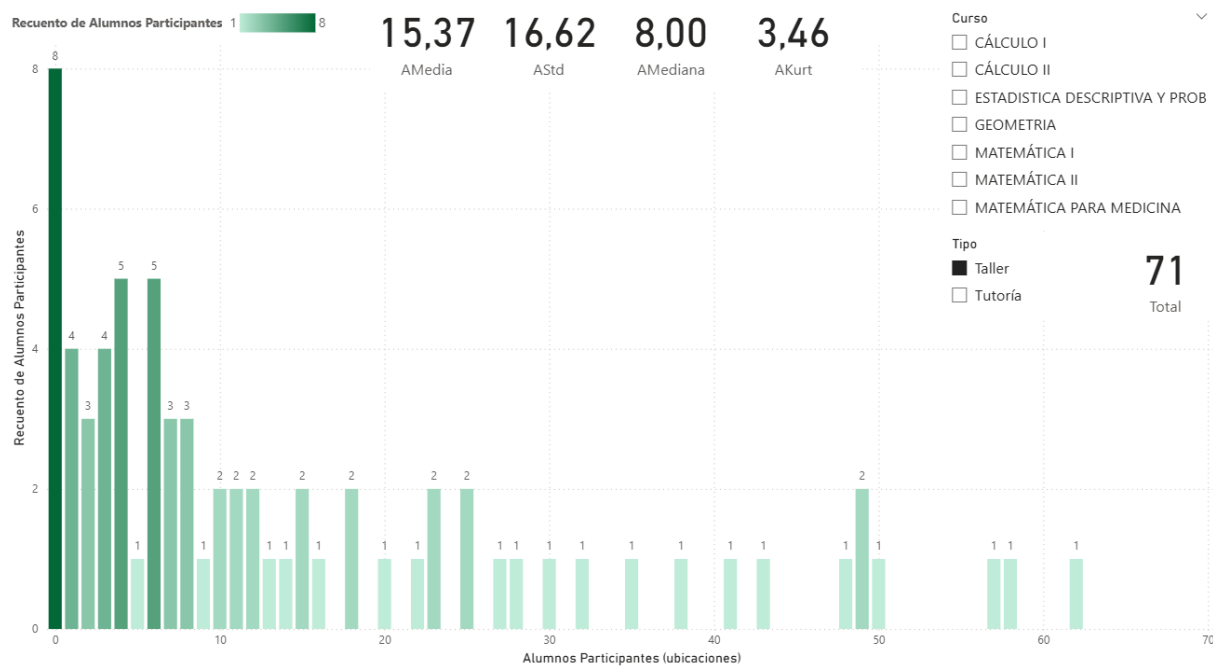


Figura 4.10: Histograma de participantes en talleres (POWER BI)

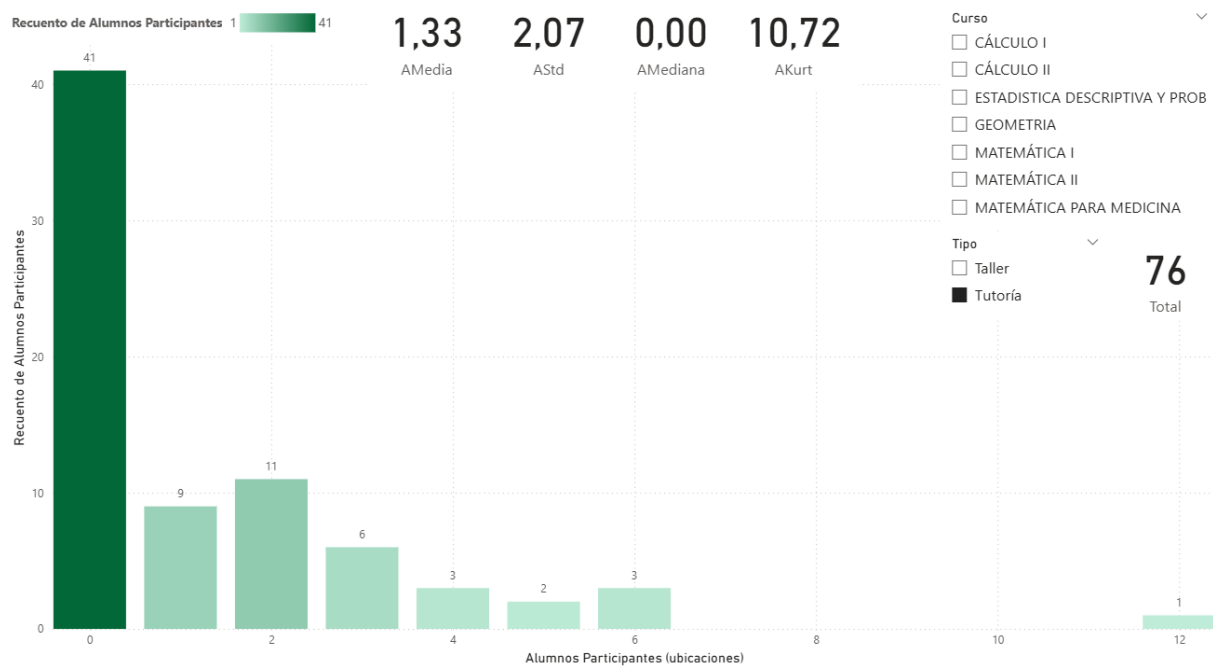


Figura 4.11: Histograma de participantes en tutorías (POWER BI)

4.3.3. Inscritos-Participantes

Algo más interesante resulta mostrar el gráfico de dispersión de $(\mathcal{I}_s, \mathcal{A}_s)$ para cada $s \in \mathcal{S}$ y visualizar la correlación de I_s y A_s para cada $s \in \mathcal{S}$.

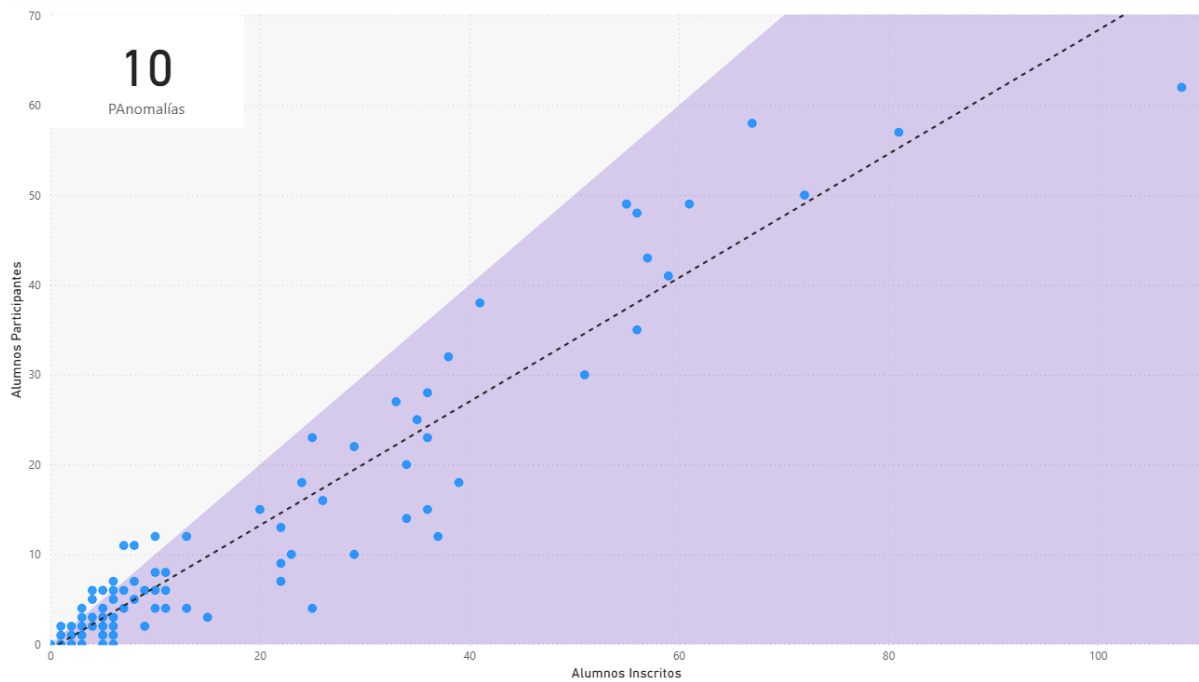


Figura 4.12: Dispersión de $(\mathcal{I}_s, \mathcal{A}_s)$ para cada $s \in \mathcal{S}$ (POWER BI)

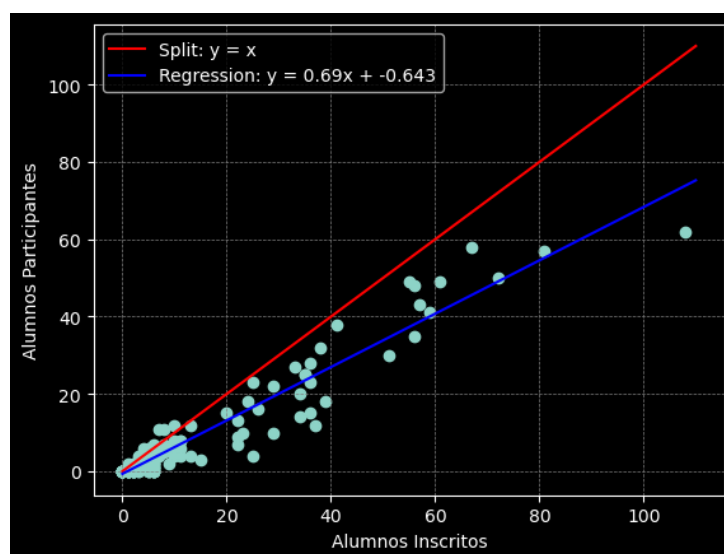


Figura 4.13: Regresión lineal de los $(\mathcal{I}_s, \mathcal{A}_s)$ (PYTHON)

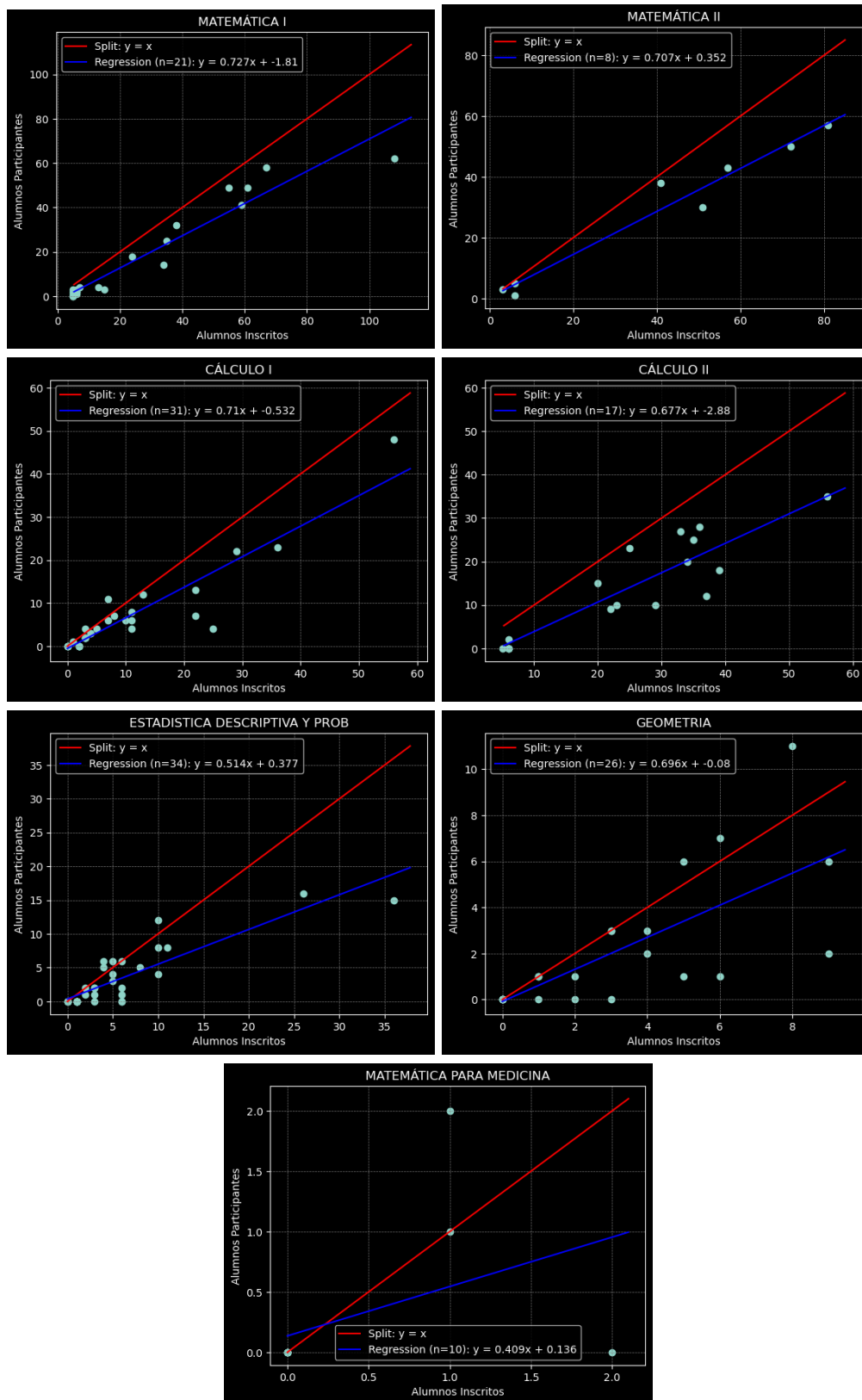


Figura 4.14: Regresión lineal de los $(\mathcal{I}_s, \mathcal{A}_s)$ por curso (PYTHON)



4.4. Duración de sesiones

La columna **Duración (minutos)**, o de forma simplificada **Duración**, no se encuentra originalmente en *MeetingsClean* sino que se añade a la misma apareciendo en *MeetingsCleanInfo*.

Esta columna contiene la diferencia (en minutos) de la hora fin la hora de inicio de la sesión en su correspondiente fila, por lo que podemos decir que

$$\text{Duración} := \text{Hora Fin} - \text{Hora Inicio}$$

Además, como esta variable es de tipo numérica continua consideramos lo siguiente al realizar la gráficas:

- Se agrupan los valores de I_s en intervalos de 8 min.
- Se muestran los estadísticos de la media, desviación estándar, mediana y kurtosis.

Más adelante se explica el motivo de la elección de estas consideraciones.

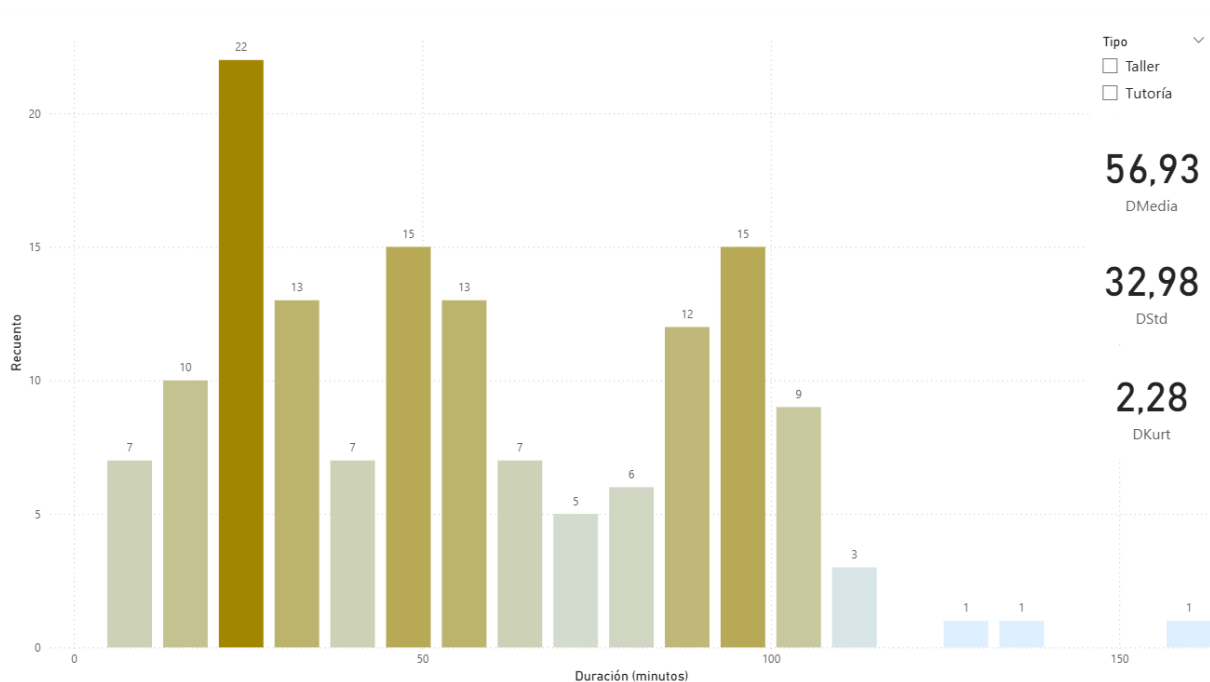


Figura 4.15: Histograma de la duración de las sesiones (POWER BI)

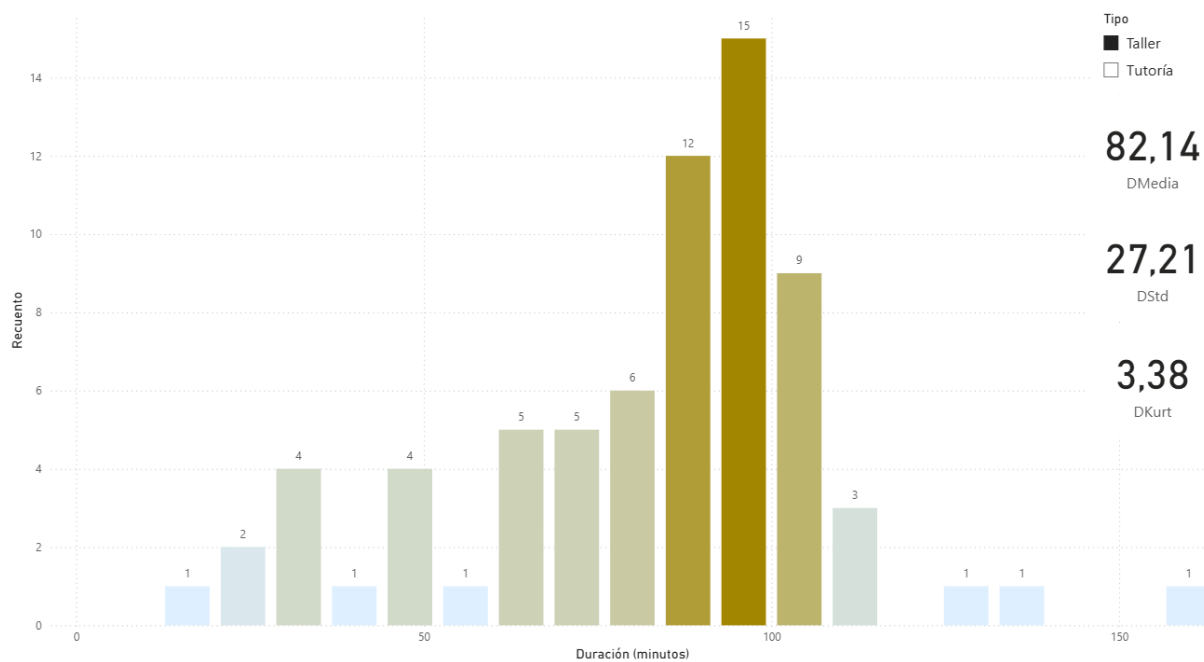


Figura 4.16: Histograma de la duración de las sesiones en talleres (POWER BI)

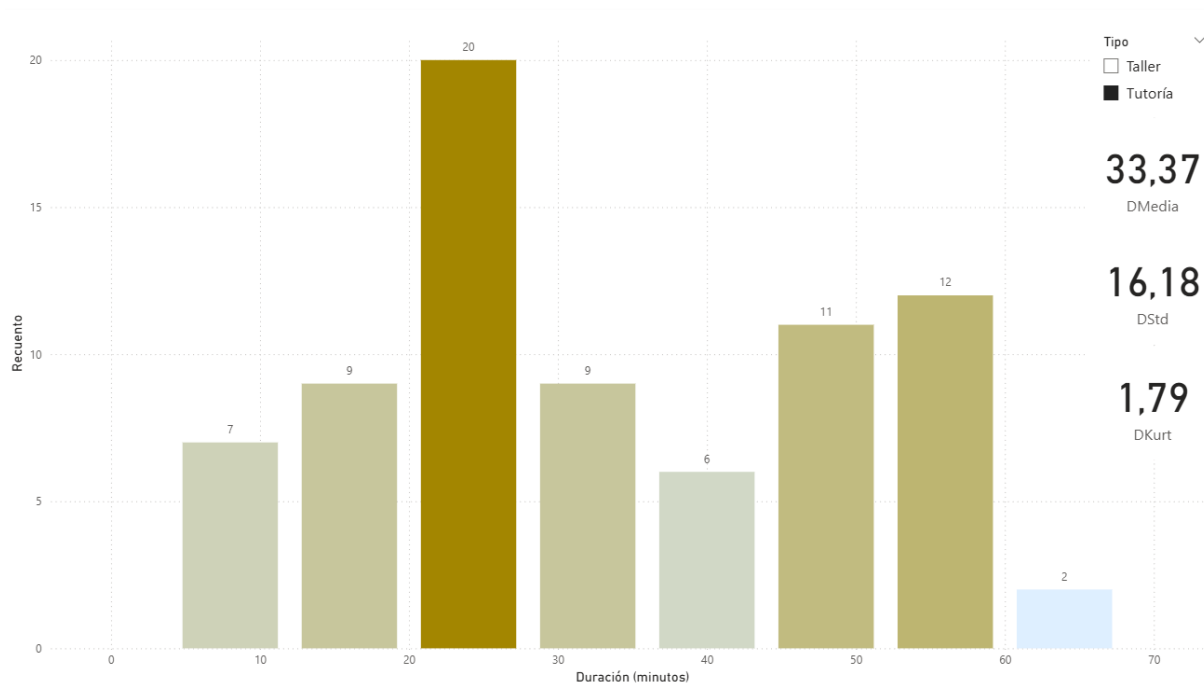


Figura 4.17: Histograma de la duración de las sesiones en tutorías (POWER BI)



4.5. Puntualidad

La columna **Puntualidad (minutos)**, o de forma simplificada **Puntualidad**, no se encuentra originalmente en *MeetingsClean* sino que se añade a la misma apareciendo en *MeetingsCleanInfo*.

Esta columna contiene la diferencia (en minutos) de la hora programada y la hora de inicio de la sesión en su correspondiente fila, por lo que podemos decir que

$$\text{Puntualidad} := \text{Hora} - \text{Hora Inicio}$$

Además, como esta variable es de tipo numérica continua consideramos lo siguiente al realizar la gráficas:

- Se agrupan los valores de **Puntualidad** en intervalos de 1 min.
- Se muestran los estadísticos de la media, desviación estándar, mediana y kurtosis.

Más adelante se explica el motivo de la elección de estas consideraciones.

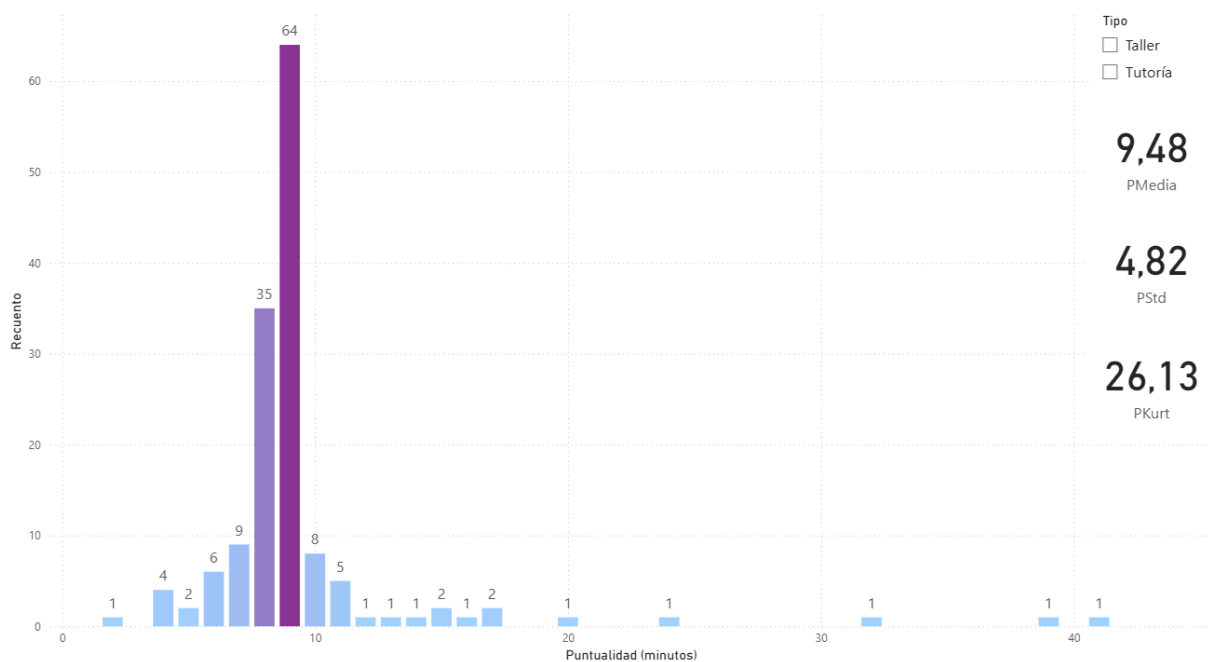


Figura 4.18: Histograma de la puntualidad (POWER BI)

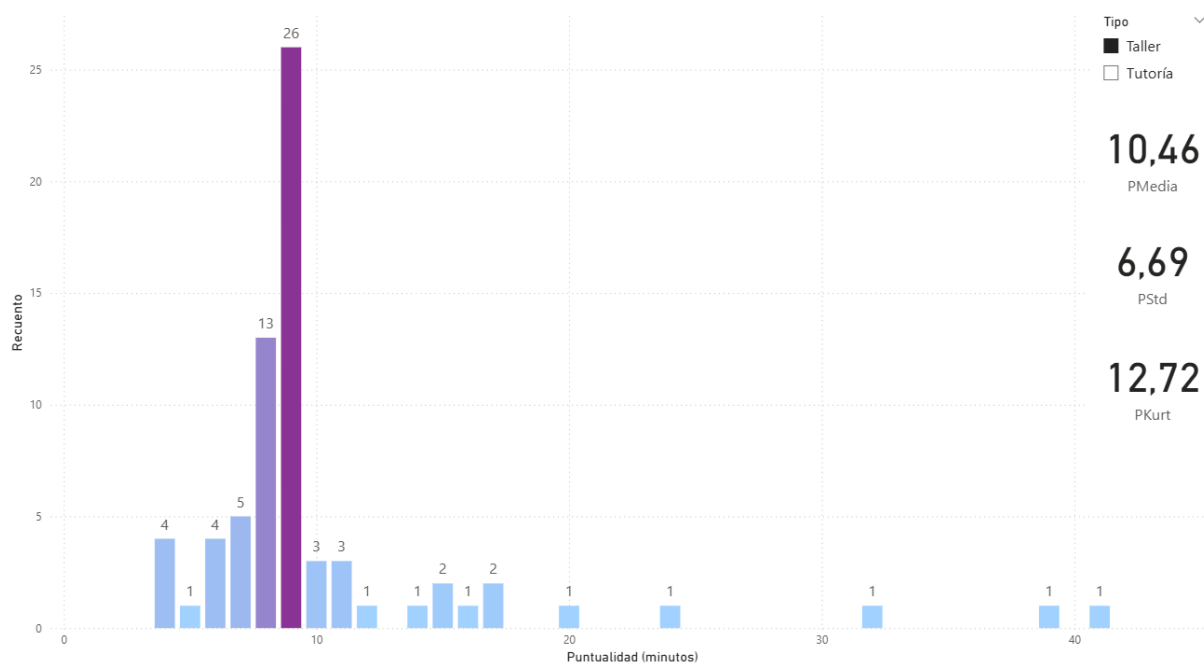


Figura 4.19: Histograma de la puntualidad en talleres (POWER BI)

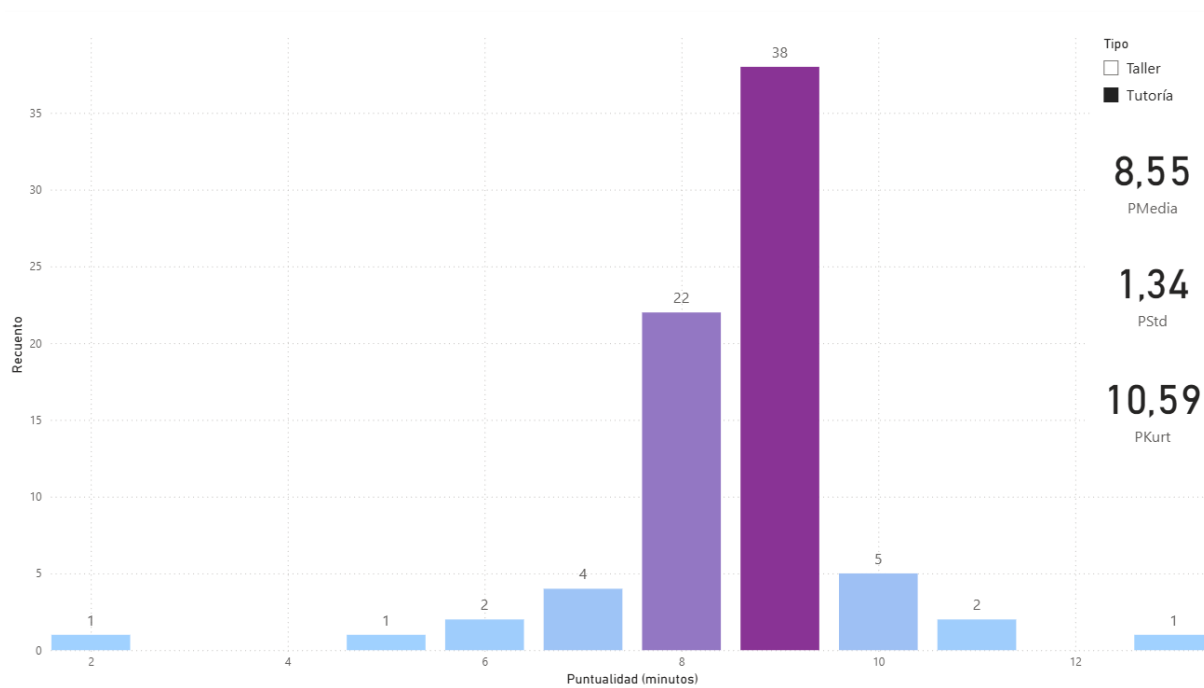


Figura 4.20: Histograma de la puntualidad en tutorías (POWER BI)



4.6. Extra: Correlación entre variables numéricas

Un gráfico interesante en el que muestra los distintos gráficos de comparación entre pares de variables numéricas de una tabla es el *pairplot* (PYTHON), que resulta ser una matriz de gráficos en cuya diagonal se muestran histogramas y en el resto se muestran gráficos de dispersión. A continuación se muestra el *pairplot* de las variables numéricas de *MeetingsCleanInfo*.

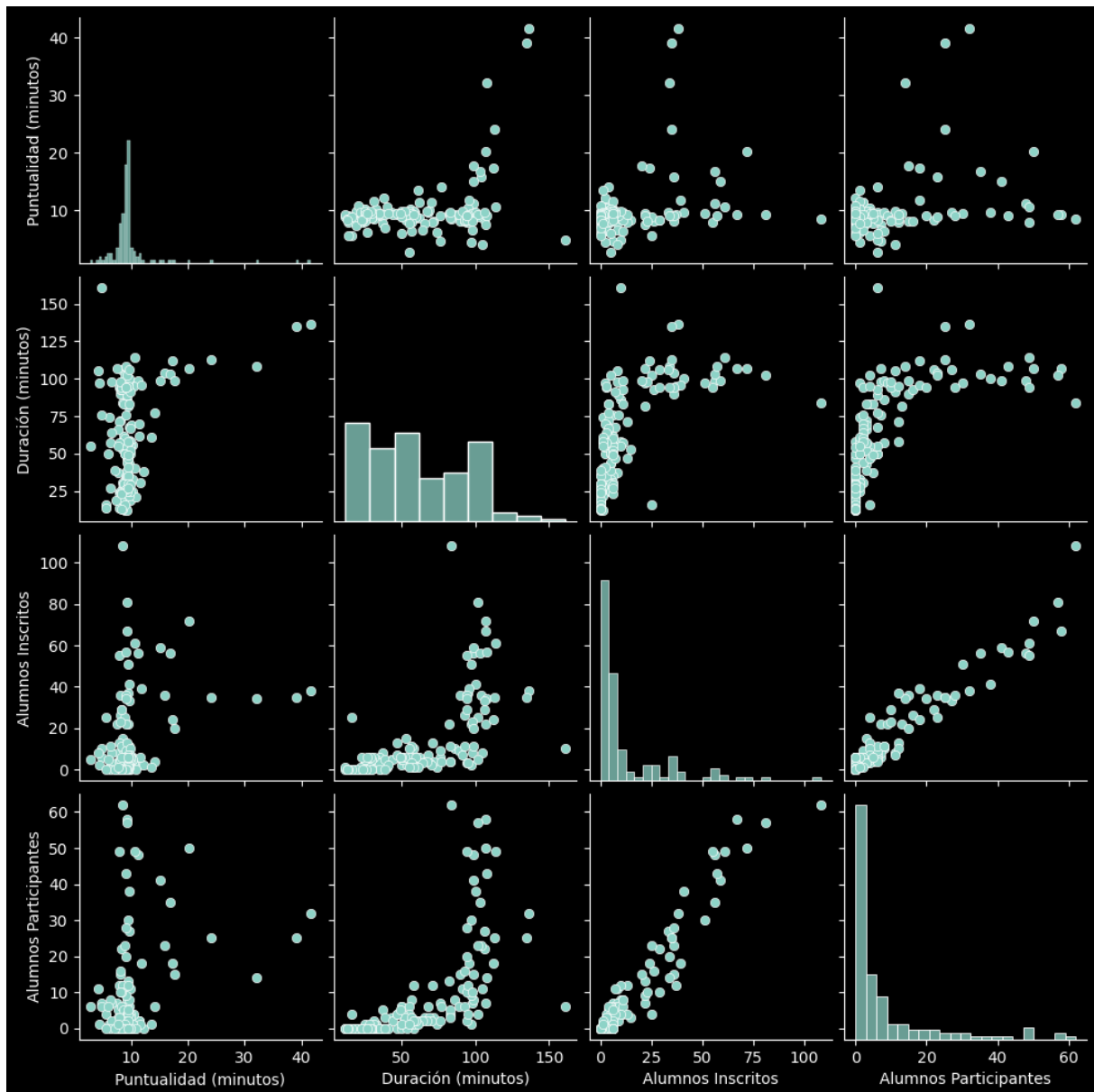


Figura 4.21: Pairplot de las variables numéricas (PYTHON)

Capítulo 5

Discusión

5.1. Tipo

Aquí la variable aleatoria analizada es T_S . Esta presenta dos posibles valores: Taller o Tutoría, por lo que se trata de una variable categórica. Al existir dos posibles valores, se esperaría que la variable aleatoria T_S sea uniformemente distribuida, es decir,

$$\mathbb{P}(T_S = \text{Taller}) = \mathbb{P}(T_S = \text{Tutoría}) = \frac{1}{2} \quad (5.1)$$

De hecho, se puede comprobar ello mediante los datos obtenidos. El siguiente gráfico muestra un gráfico de barras obtenida a partir de *MeetingsCleanInfo* que resume la aproximación obtenida de la distribución de las variables aleatorias T_S .

Con ello formamos la tabla

t_s	$\hat{\mathbb{P}}_{T_S}(t_s)$
Taller	$\frac{71}{147}$
Tutoría	$\frac{76}{147}$

Tabla 5.1: Aproximación de la distribución de T_S

Entonces, de acuerdo a lo planteado en 5.1 y usando las aproximaciones planteadas en 2.3 se puede comprobar que la aproximación obtenida de la distribución de T_S es correcta pues de acuerdo a la gráfica de la Figura 4.1.

$$\hat{\mathbb{P}}_{T_S}(\text{Taller}) = \frac{76}{147} \approx 0.517, \quad \hat{\mathbb{P}}_{T_S}(\text{Tutoría}) = \frac{71}{147} \approx 0.483 \quad (5.2)$$

Ahora, veamos qué sucede con la distribución de T_S para cada uno de los cursos, es decir, T_S dado C .

Dado que tratamos con pocos cursos, es decir $n(\mathcal{C}) = 7$, podemos mostrar en una tabla la aproximación obtenida de la distribución de T_S dado C .

$\hat{\mathbb{P}}_{C=c}(T_S = t_s)$		t_s	
		Taller	Tutoría
c	Matemática I	$\frac{9}{21}$	$\frac{12}{21}$
	Matemática II	$\frac{5}{8}$	$\frac{3}{8}$
	Cálculo I	$\frac{22}{31}$	$\frac{9}{31}$
	Cálculo II	$\frac{12}{17}$	$\frac{5}{17}$
	Estadística Descriptiva	$\frac{10}{34}$	$\frac{24}{34}$
	Geometría	$\frac{10}{26}$	$\frac{16}{26}$
	Matemática para Medicina	$\frac{3}{10}$	$\frac{7}{10}$

Tabla 5.2: Aproximación de la distribución de T_S dado C

5.2. Curso

Aquí la variable aleatoria analizada es C , del cual conocemos los posibles valores que toma ya que conocemos los elementos de \mathcal{C} como se muestra en 3.3. Por distintos motivos (como demanda) es muy posible que existan cursos con mayor o menor asignación, por lo que no se puede intuir la distribución de C .

En la Figura 4.3 muestra un gráfico de barras obtenido a partir de *MeetingsCleanInfo* que resume la aproximación obtenida de la distribución de la variable aleatoria C . Con ello formamos la tabla

c	$\hat{\mathbb{P}}_{\mathcal{C}}(c)$
Matemática I	$\frac{21}{147}$
Matemática II	$\frac{8}{147}$
Cálculo I	$\frac{31}{147}$
Cálculo II	$\frac{17}{147}$
Estadística Descriptiva	$\frac{34}{147}$
Geometría	$\frac{26}{147}$
Matemática para Medicina	$\frac{10}{147}$

Tabla 5.3: Aproximación de la distribución de C

En base a la Figura 4.2, mostramos la tabla de la aproximación de la distribución de C dado T_S . Se resalta el hecho de que $n(\mathcal{S} \cap \{s \mid t_s = \text{Taller}\}) = 76$ y $n(\mathcal{S} \cap \{s \mid t_s = \text{Tutoría}\}) = 71$ ya que

5.3. INSCRITOS Y PARTICIPANTES

estos valores aparecen en la parte superior de la Figura 4.1.

$\hat{\mathbb{P}}_{T_S=t_s}(C=c)$		t_s	
		Taller	Tutoría
c	Matemática I	$\frac{9}{71}$	$\frac{12}{76}$
	Matemática II	$\frac{5}{71}$	$\frac{3}{76}$
	Cálculo I	$\frac{22}{71}$	$\frac{9}{76}$
	Cálculo II	$\frac{12}{71}$	$\frac{5}{76}$
	Estadística Descriptiva	$\frac{10}{71}$	$\frac{24}{76}$
	Geometría	$\frac{10}{71}$	$\frac{16}{76}$
	Matemática para Medicina	$\frac{3}{71}$	$\frac{7}{76}$

Tabla 5.4: Aproximación de la distribución de C dado T_S

5.3. Inscritos y participantes

5.3.1. Inscritos

Como sabemos, \mathcal{I} representa la variable aleatoria número de alumnos inscritos en una sesión y mediante 2.2 se puede obtener su distribución aproximada, o también mediante 2.5.

Incluso si los talleres solo permiten un máximo de 100 alumnos, en la Figura 4.6 se observa que existe un único valor dentro de *MeetingsCleanInfo* con $I_s > 100$. El fenómeno no podría explicarse a detalle pero se hace el supuesto de que se sucede debido al cómo la plataforma donde los alumnos realizan sus inscripciones registra los datos. Sin embargo, al tratarse de un único valor este no es tan relevante como para ser excluido de la muestra ya que podemos ‘ignorar’ que el límite de 100 inscripciones existe.

Lo que es bastante claro de apreciar es que los valores de I_s más frecuentes tienden a ser los que tienen valores más pequeños pero con una alta dispersión. Es decir, el conteo de alumnos inscritos es bastante concentrado en un pequeño rango de valores.

5.3.2. Participantes

Como sabemos, \mathcal{A} representa la variable aleatoria número de alumnos participantes o asistentes en una sesión y mediante 2.2 se puede obtener su distribución aproximada, o también mediante 2.5. En este caso aparentemente no se observan valores atípicos visibles en la Figura 4.9 (a

diferencia de lo mostrado en la Figura 4.6). Lo que sí es apreciable es que la tendencia en la distribución de A_s es bastante similar a la de I_s .

5.3.3. Inscritos-Participantes

La relación que tienen \mathcal{I} y \mathcal{A} parece tener algún tipo de dependencia. Como es usual con datos numéricos, se puede analizar los coeficientes de correlación (Pearson) entre los mismos. Mediante el uso de librerías de PYTHON se puede obtener la matriz de correlación (Pearson) asociada.

	Puntualidad (minutos)	Duración (minutos)	Alumnos Inscritos	Alumnos Participantes
Puntualidad (minutos)	1.000000	0.363871	0.310004	0.307390
Duración (minutos)	0.363871	1.000000	0.635021	0.654129
Alumnos Inscritos	0.310004	0.635021	1.000000	0.959499
Alumnos Participantes	0.307390	0.654129	0.959499	1.000000

Figura 5.1: Correlación de las variables numéricas

Lo que muestra la matriz de correlación permitiría explicar el por qué de la similitud entre los histogramas de las variables aleatorias I e A . El coeficiente de correlación entre ambas variables es muy próxima a 1 lo que indica una alta correlación positiva, y además es el coeficiente de correlación más alto de todas las variables numéricas.

En la Figura 4.12 se muestra la dispersión de $(\mathcal{I}_s, \mathcal{A}_s)$ para $s \in \mathcal{S}$, es decir, los pares inscritos-participantes obtenidos a partir de *MeetingsCleanInfo*. La región sombreada representa la región de pares inscritos-participantes que tiene un comportamiento ‘regular’, es decir, que $\mathcal{I}_s \geq \mathcal{A}_s$ (ya que no es posible que alguien asista sin estar previamente inscrito, en teoría). En base a ello, sólo se contabilizan 10 pares que no presentan el comportamiento ‘regular’ y los consideramos como atípicos o ‘anómalos’.

Por otro lado, dado que la correlación lineal entre \mathcal{I} y \mathcal{A} es bastante alta, podemos usar la regresión lineal (bajo el modelo $\mathcal{A} \approx \beta\mathcal{I} + \alpha + \varepsilon$) para obtener una aproximación de la distribución de \mathcal{A} dado \mathcal{I} . En la Figura 4.12 se muestra la recta de regresión lineal en base a los datos obtenidos y aparece en forma de línea punteada. Sin embargo, como este gráfico fue elaborado en POWER BI, este no se muestra con una ecuación explícita pero podemos encontrarla utilizando las librerías existentes en PYTHON.

Vamos a centrarnos en la ecuación de la recta de regresión lineal obtenida en la Figura 4.13. Entonces para $x \in \mathbb{R}^+ \cap \mathbb{Z}$ representando el número de alumnos inscritos en una sesión (que es conocido antes de una sesión) tendríamos que $y = 0.69x - 0.643$, con un redondeo adecuado, representaría el número estimado de asistentes en la sesión.



Si bien esto muestra el comportamiento global del par inscritos-participantes se desconoce el comportamiento según el curso.

Por otro lado, si trabajamos en base al supuesto $\mathcal{A} \approx 0.69\mathcal{I} - 0.643 + \varepsilon_{\mathcal{I}}$ con $\varepsilon_{\mathcal{I}} \sim \mathcal{N}(0, \sigma_{\mathcal{I}}^2)$ entonces

$$\mathbb{E}[\mathcal{A} \mid \mathcal{I} = n] \approx 0.69n - 0.643$$

y deberíamos ser capaces de verificar que $\mathcal{A} - 0.69\mathcal{I} + 0.643 \sim \varepsilon_{\mathcal{I}} \sim \mathcal{N}(0, \sigma_{\mathcal{I}}^2)$ mediante el contraste de hipótesis de Shapiro-Wilk (Ejemplo 7.3.8). De hecho, realizando el procedimiento de contraste de hipótesis de Shapiro-Wilk sobre $\varepsilon_{\mathcal{I}_s}$ para cada \mathcal{I}_s .

	Inscritos	size	statistic	p_value	H_o
0	0	25	1.0000	1.0000	Accepted
1	1	13	0.7092	0.0007	Rejected
2	2	9	0.6843	0.0009	Rejected
3	3	12	0.9098	0.2123	Accepted
4	4	6	0.8663	0.2117	Accepted
5	5	12	0.9009	0.1630	Accepted
6	6	17	0.8183	0.0037	Rejected
7	7	3	0.9423	0.5367	Accepted
8	8	3	0.9643	0.6369	Accepted
9	10	4	0.9714	0.8500	Accepted
10	11	4	0.8634	0.2725	Accepted
11	22	3	0.9643	0.6369	Accepted
12	36	3	0.9826	0.7470	Accepted

Figura 5.2: Contraste de hipótesis de Shapiro-Wilk sobre $\varepsilon_{\mathcal{I}_s}$ para cada \mathcal{I}_s (PYTHON)

Recordemos que $H_0 : \varepsilon_{\mathcal{I}}$ es una variable aleatoria normal. En la Figura 5.2 se muestra que para la mayoría de los \mathcal{I}_s , la hipótesis nula es **aceptada**, por lo que en promedio $\varepsilon_{\mathcal{I}}$ es una variable aleatoria normal. Además, debemos tener en consideración que $n(\mathcal{S})$ no es muy grande como para contar con suficientes datos para cada \mathcal{I}_s .

Por otro lado, al observar las ecuaciones de las rectas de regresión en la Figura 4.14 (bajo el modelo $\mathcal{A} \approx \beta_i \mathcal{I} + \alpha_i + \varepsilon_{\mathcal{I}}$ para cada $i \in C$) podemos ver que estos comportamientos son bastante similares, principalmente en el sentido de la pendiente. Si promediamos las pendientes β_i de las rectas obtenidas por curso obtenemos

$$\hat{\beta} = \frac{1}{7} \sum_{i \in C} \beta_i = \frac{0.727 + 0.707 + 0.710 + 0.677 + 0.514 + 0.696 + 0.409}{7} = \frac{4.44}{7} \approx 0.634$$

y el error relativo de la pendiente β a $\hat{\beta}$ de la muestra total es 8.07 %, lo cual es pequeño. Esto podría indicar que, de manera aproximada, existe una independencia entre las variables aleatorias $(\mathcal{I}, \mathcal{A})$ y C .

Observación 5.3.1. *Dado que \mathcal{I} y \mathcal{A} presentan múltiples combinaciones de n y a para evaluar la estimación de $\mathbb{P}(\mathcal{I} = n \mid \mathcal{A} = a)$, se descarta mostrar una tabla que muestre explícitamente las aproximaciones. Por lo tanto, nos contentamos con el estudio de la correlación y la normalidad. Sin embargo, esta tabla se puede encontrar en el repositorio de GitHub [One25] en el archivo .ipynb.*

5.4. Duración de las sesiones

El motivo de agrupar en intervalos de 8 min es porque si buscamos partir en 9 subintervalos los intervalos $[0; 45]$ y $[0; 90]$, los intervalos tendrían longitud 5 min y 10 min respectivamente, por lo que para evitar el tener que utilizar intervalos de longitud diferentes se toman intervalos que resulten del promedio aproximado de los mismos que es $7.5 \text{ min} \approx 8 \text{ min}$.

En el comportamiento global de esta variable podría no ser apreciable algún comportamiento en particular. Sin embargo, como ya se ha mencionado, según el tipo de sesión la duración máxima establecida es: 90 min para los talleres y 45 min para las tutorías. Por lo tanto, se espera que la mayor parte de los valores registrados en la columna **Duración (minutos)** estén mayoritariamente concentrados en 90 y 45 para los talleres y tutorías, respectivamente.

Lo que es apreciable es que en talleres la duración de las sesiones suele ser la máxima establecida (90 min), mientras que en tutorías suele ser la mitad de su máxima establecida (22.5 min). Esto es debido a que la permanencia del tutor a cargo (mi persona) es hasta un máximo aproximado de la mitad del tiempo cuando no ingresan estudiantes a la sesión, por lo que esto también indicaría que las sesiones de tutorías son las que menos se dictan en comparación con los talleres.

5.5. Puntualidad

Dado que el reglamento de la *Universidad Tecnológica del Perú* indica que, tanto para talleres como para tutorías, la hora de ingreso a la sesión debe ser entre 5 y 10 minutos antes de la hora programada, entonces se esperaría que los valores registrados en la columna **Puntualidad (minutos)** se encuentren mayoritariamente concentrados en el intervalo $[5; 10]$.

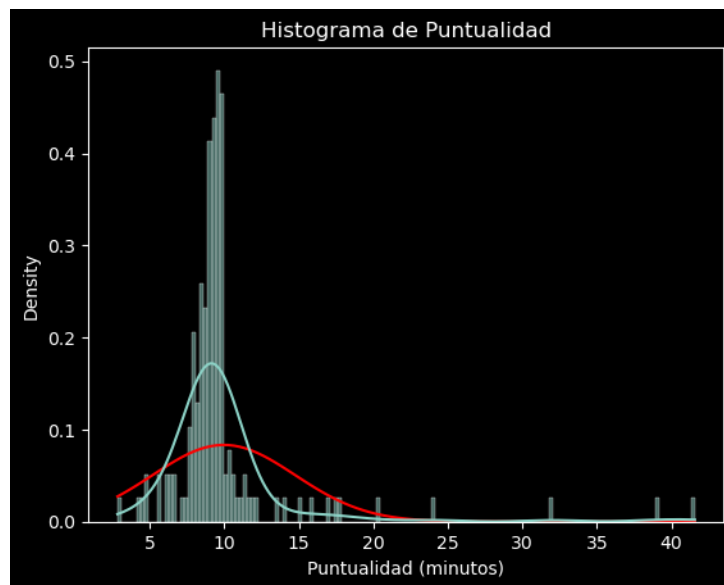


Figura 5.3: Histograma de la puntualidad (PYTHON).

La Figura 5.3 muestra un histograma de la variable aleatoria **Puntualidad** con curvas que estiman la distribución. En la línea en rojo muestra la función de densidad de una variable aleatoria normal con media y desviación estándar de la muestra en **Puntualidad** (Figura 4.18). Por otro lado, en la línea en celeste muestra la distribución aproximada usando *KDE* (Kernel Density Estimation).

Sin embargo, también debemos tener en cuenta que podrían presentarse valores atípicos, pues como ya se mencionó *Zoom* registra que una sesión ha sido iniciada si es que existe algún integrante en la sesión e independiente de si se trata del anfitrión (mi persona) o no.

En efecto, los valores atípicos se aparecen y estos son los valores de **Puntualidad** más altos. Esto podría explicarse mediante el hecho de que algún estudiante ingresó a la sesión minutos muy antes de la hora programada y permaneció hasta la hora en que el anfitrión ingresó.

Además, el comportamiento global es muy similar al observado al comportamiento en talleres y tutorías. Esto tiene sentido, pues, debido a que la puntualidad es indistinta al tipo de sesión.

Capítulo 6

Conclusiones

1. Se asignan equitativamente sesiones de tutoría y taller.
2. La cantidad de inscritos y participantes en talleres es mayor a la cantidad de inscritos y participantes en tutorías.
3. El número de asistentes presenta una correlación lineal positiva fuerte con el número de inscritos.
4. Se espera que de una cantidad de alumnos inscritos en cualquier tipo o curso, aproximadamente el 70 % de ellos asistan a la misma. Más específicamente, se espera que de cierta cantidad de inscritos a una sesión
 - a) en el curso de Matemática I aproximadamente el 72.7 % de ellos asistan a la misma;
 - b) en el curso de Matemática II aproximadamente el 70.7 % de ellos asistan a la misma;
 - c) en el curso de Cálculo I aproximadamente el 71.0 % de ellos asistan a la misma;
 - d) en el curso de Cálculo II aproximadamente el 67.7 % de ellos asistan a la misma;
 - e) en el curso de Estadística Descriptiva aproximadamente el 51.4 % de ellos asistan a la misma;
 - f) en el curso de Geometría aproximadamente el 69.6 % de ellos asistan a la misma;
 - g) en el curso de Matemática para Medicina aproximadamente el 40.9 % de ellos asistan a la misma.
5. La duración de las sesiones en talleres suele ser la máxima establecida, mientras que en las tutorías suele ser la media.
6. La puntualidad mantiene su distribución aproximadamente igual en talleres y tutorías.

Capítulo 7

Anexos

7.1. Miscelánea

Definición 7.1.1 (Función indicatriz). Sea A un subconjunto de un conjunto X , definimos la función indicatriz de A sobre X como $\mathbb{1}_A : X \longrightarrow \{0, 1\}$ con

$$\mathbb{1}_A(x) = \begin{cases} 1 & , \ x \in A \\ 0 & , \ x \in A^c \end{cases}$$

7.2. Probabilidad

Definición 7.2.1 (Probabilidad condicional). Sean A y B eventos de un espacio de probabilidad $(\Omega, \mathcal{F}, \mathbb{P})$. La probabilidad de A dado B se define como

$$\mathbb{P}(A \mid B) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \quad (7.1)$$

Definición 7.2.2 (Partición). Decimos que los $A_1, A_2, \dots, A_n \subset A$ forman una partición de A si los A_i son exhaustivos y mutuamente excluyentes.

Definición 7.2.3 (Partición de un espacio de probabilidad). Decimos que los $A_1, A_2, \dots, A_n \subset A$ forman una partición de un espacio de probabilidad $(\Omega, \mathcal{F}, \mathbb{P})$ si $A_i \in \mathcal{F}$, $i = 1, \dots, n$ y forman una partición de Ω .

Teorema 7.2.4 (Bayes). Sea A_1, A_2, \dots, A_n una partición del espacio de probabilidad $(\Omega, \mathcal{F}, \mathbb{P})$ tales que $\mathbb{P}(A_i) > 0$, $i = 1, \dots, n$. Sea B un evento arbitrario de $(\Omega, \mathcal{F}, \mathbb{P})$, entonces

$$\mathbb{P}(A_i \mid B) = \frac{\mathbb{P}(B \mid A_i) \mathbb{P}(A_i)}{\mathbb{P}(B)} \quad (7.2)$$

Teorema 7.2.5 (Probabilidad total). *Bajo las mismas condiciones del Teorema 7.2.4,*

$$\mathbb{P}(B) = \sum_{i=1}^n \mathbb{P}(B \mid A_i) \mathbb{P}(A_i) \quad (7.3)$$

Además se obtiene la **fórmula de Bayes**

$$\mathbb{P}(A_i \mid B) = \frac{\mathbb{P}(B \mid A_i) \mathbb{P}(A_i)}{\sum_{i=1}^n \mathbb{P}(B \mid A_i) \mathbb{P}(A_i)} \quad (7.4)$$

Definición 7.2.6 (Distribución normal). Una variable aleatoria X es normalmente distribuida con media μ y varianza σ^2 si su densidad es

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \quad (7.5)$$

y en tal caso denotamos $X \sim \mathcal{N}(\mu, \sigma^2)$. Además, si $\mu = 0$ y $\sigma = 1$ se dice que X es normalmente distribuida.

Teorema 7.2.7 (Ley débil de los grandes números). *Sea X_1, X_2, X_3, \dots una sucesión de variables aleatorias independientes e idénticamente distribuidas con valor esperado μ y varianza σ^2 , entonces el promedio*

$$\bar{X}_n = \frac{X_1 + \dots + X_n}{n} = \frac{1}{n} \sum_{i=1}^n X_i$$

converge en probabilidad a μ . En otras palabras, para cualquier $\varepsilon > 0$ se cumple que

$$\lim_{n \rightarrow \infty} P(|\bar{X}_n - \mu| < \varepsilon) = 1.$$

Teorema 7.2.8 (Ley fuerte de los grandes números). *Sea X_1, X_2, X_3, \dots una sucesión de variables aleatorias independientes e idénticamente distribuidas que cumplen $\mathbb{E}[X_i] < \infty$ y tienen valor esperado $\mathbb{E}[X_i] = \mu$, entonces*

$$\mathbb{P}\left(\lim_{n \rightarrow \infty} \bar{X}_n = \mu\right) = 1,$$

es decir, el promedio de las variables aleatorias converge a μ casi seguramente.

7.3. Métricas estadísticas

Definición 7.3.1 (Matriz de covarianza). Sea $X = (X_1, X_2, \dots, X_n)$ una variable aleatoria en \mathbb{R}^n con X_i de varianza y media finitas. La **matriz de covarianza** de X se define como

$$K_{XX} := \text{Var}(X) := \text{cov}(X, X) := \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^T]$$



Definición 7.3.2 (Estimador (estadístico)). Es una función generada a partir de los datos de una muestra que se usa para estimar algún parámetro desconocido de la población. Cuando el estimador toma un valor en particular en base a los datos de una muestra, se llama **estimador puntual**.

Definición 7.3.3 (Sesgo de un estimador). Sea $\hat{\theta}$ un estimador de un parámetro θ . Entonces el sesgo o *bias* de $\hat{\theta}$ se define como

$$B(\hat{\theta}) = \mathbb{E} [\hat{\theta}] - \theta$$

Si $B(\hat{\theta}) = 0$, entonces $\hat{\theta}$ decimos que es un **estimador insesgado**.

Definición 7.3.4 (Consistencia de un estimador). Sea $\hat{\theta}_n$ un estimador de un parámetro θ determinado por una muestra de tamaño n . Decimos que $\hat{\theta}_n$ es consistente si

$$\lim_{n \rightarrow \infty} \mathbb{P} (\hat{\theta}_n = \theta) = 1$$

Definición 7.3.5 (Nivel de significación). Sea un estadístico s en un contraste de hipótesis con hipótesis nula H_0 . En general, decimos que $\alpha \in [0; 1]$ es un nivel de significación si siendo p_s el p -valor del estadístico s , entonces:

- Se acepta H_0 si $p_s \geq \alpha$.
- Se rechaza H_0 si $p_s < \alpha$.

En particular, sea Z normal estándar, decimos que $\alpha \in [0; 1]$ es un nivel de significación en los siguientes casos:

- **Cola izquierda:** $\alpha = \mathbb{P} (Z \leq z_\alpha)$ o $1 - \alpha = \mathbb{P} (Z > z_\alpha)$. Si
 - $s \geq z_\alpha$ entonces se acepta H_0 .
 - $s < z_\alpha$ entonces se rechaza H_0 .
- **Cola derecha:** $\alpha = \mathbb{P} (Z > z_\alpha)$ o $1 - \alpha = \mathbb{P} (Z \leq z_\alpha)$.
 - $s \leq z_\alpha$ entonces se acepta H_0 .
 - $s > z_\alpha$ entonces se rechaza H_0 .
- **Dos colas:** $\alpha = \mathbb{P} (|Z| > z_{\frac{\alpha}{2}})$ o $1 - \alpha = \mathbb{P} (|Z| \leq z_{\frac{\alpha}{2}})$.
 - $|s| \leq z_{\frac{\alpha}{2}}$ entonces se acepta H_0 .
 - $|s| > z_{\frac{\alpha}{2}}$ entonces se rechaza H_0 .

En todos los casos z_α es el cuantil normal de nivel $1 - \alpha$. Cuando $\alpha = 0.05$, entonces $z_{0.05} \approx -1.65$ en la cola izquierda y $z_{0.05} \approx 1.65$ en la cola derecha.

Definición 7.3.6 (Contraste de hipótesis). Es un método estadístico que permite decidir si un conjunto de datos provee de suficiente evidencia para rechazar o no una hipótesis. La hipótesis puede ser de dos tipos: hipótesis nula (H_0) e hipótesis alternativa (H_1). La hipótesis nula es la que se aceptará o rechazará según algún tipo de test estadístico.

Ejemplo 7.3.7 (Contraste de la proporción (dos colas)). Sean las hipótesis:

$$H_0 : p = p_0$$

$$H_1 : p \neq p_0$$

donde p es la probabilidad de éxito y $q = 1 - p$ es la probabilidad de fracaso y p_0 es el valor supuesto para la probabilidad de éxito en la hipótesis nula.

Para un nivel de significación α , es necesario determinar el valor del cuantil $z_{\alpha/2}$ de una distribución normal estándar. Para la proporción muestral, el estadístico de contraste viene dado por:

$$z_c = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 \cdot q_0}{n}}},$$

donde \hat{p} es la proporción muestral, $q_0 = 1 - p_0$ y n el tamaño de la muestra.

Luego,

- Si $|z_c| > z_{\alpha/2}$, se rechaza H_0 .
- Si $|z_c| \leq z_{\alpha/2}$, no se rechaza (o se acepta que se tiene suficiente evidencia) H_0 .

Ejemplo 7.3.8 (Contraste de normalidad (Shapiro-Wilk)). Sea una muestra x_1, x_2, \dots, x_n y supongamos las hipótesis:

H_0 : la muestra proviene de una distribución normal

H_1 : la muestra no proviene de una distribución normal

El estadístico de contraste está dado por

$$W = \frac{(\sum_{i=1}^n a_i x_{(i)})^2}{\sum_{i=1}^n (x_i - \bar{x})^2},$$

donde:

- $x_{(i)}$ denota el i -ésimo valor más pequeño de la muestra (estadístico de orden i)
- $\bar{x} = \frac{x_1 + \dots + x_n}{n}$ es la media muestral
- Los coeficientes a_i están dados por

$$(a_1, \dots, a_n) = \frac{m^T V^{-1}}{C}, \quad C = \|V^{-1}m\| = \left(m^T V^{-1} V^{-1} m\right)^{1/2}$$

7.3. MÉTRICAS ESTADÍSTICAS

- El vector $m = (m_1, \dots, m_n)^\top$ está formado por los valores esperados de los estadísticos de orden de las variables aleatorias independientes e idénticamente distribuidas, muestreadas de una distribución normal estándar.
- V es la matriz de covarianza los estadísticos de orden normales, es decir,

$$V = \text{cov}(x_{(1)}, \dots, x_{(n)})$$

Luego, para un nivel de significación α , si p_W es el p -valor del estadístico W , entonces

- Si $p_W < \alpha$, se rechaza H_0 .
- Si $p_W \geq \alpha$, no se rechaza (o se acepta que se tiene suficiente evidencia) H_0 .

Definición 7.3.9 (Kurtosis). Sea X una variable aleatoria con media μ y varianza σ^2 . La **kurtosis** de X se define como

$$\kappa(X) := \mathbb{E} \left[\frac{(X - \mu)^4}{\sigma^4} \right] = \frac{\mathbb{E}[(X - \mu)^4]}{\mathbb{E}[(X - \mu)^2]^2}$$

Además, el exceso de kurtosis de X se define como $\kappa_e(X) = \kappa(X) - 3$ y

- Si $\kappa_e(X) > 0$, entonces X tiene una distribución **mesokúrtica**.
- Si $\kappa_e(X) < 0$, entonces X tiene una distribución **platikúrtica**.
- Si $\kappa_e(X) = 0$, entonces X tiene una distribución **mesokúrtica**.

Definición 7.3.10 (Coeficiente de correlación (Pearson)). Sea X y Y dos variables aleatorias. Entonces el coeficiente de correlación de Pearson entre X y Y se define como

$$\rho(X, Y) := \frac{\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]}{\sqrt{\mathbb{E}[X^2] - \mathbb{E}[X]^2} \sqrt{\mathbb{E}[Y^2] - \mathbb{E}[Y]^2}}$$

Definición 7.3.11 (Matriz de correlación (Pearson)). Sea X_1, X_2, \dots, X_n una sucesión de variables aleatorias, la matriz de correlación (Pearson) de X_1, X_2, \dots, X_n se define como

$$\rho := [\rho(X_i, X_j)]_{n \times n}$$

Bibliografía

- [Gal22] J.F.L. Gall. *Measure Theory, Probability, and Stochastic Processes*. Graduate Texts in Mathematics. Springer International Publishing, 2022. ISBN: 9783031142055. URL: <https://books.google.com.pe/books?id=Ba2YEAAAQBAJ>.
- [Kle13] A. Klenke. *Probability Theory: A Comprehensive Course*. Universitext. Springer London, 2013. ISBN: 9781447153603.
- [Lim98] E.L. Lima. *Algebra Lineal*. Colección textos del IMCA. Instituto de Matemática y Ciencias Afines, UNI, 1998.
- [One25] OneDividedByX. *GitHub Repositories*. 2025. URL: <https://github.com/OneDividedByX/EDA---Registered-Attendants/tree/main/Sources/Data>.
- [Rui95] C.P. Ruiz. *Cálculo vectorial*. Prentice Hall Hispanoamericana, S.A., 1995. ISBN: 9789688805299.
- [Tor] University of Toronto. *Random Vectors and Matrices*. URL: <https://www.utstat.toronto.edu/~brunner/oldclass/appliedf11/handouts/2101f11RandomVectorsMVN.pdf>.
- [Unk12] Unknown. 2012. URL: <https://www.ugr.es/~mvargas/Infe2.pdf>.
- [UTP] UTP. *UTP Reservas*. URL: <https://reservarecursos.utp.edu.pe/ref-acad>.
- [Zoo] Zoom. *Informes*. URL: <https://zoom.us/account/report?isPersonal=true#/usageReports>.