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

26 de septiembre de 2025

Índice general

Resumen

Capítulo 1

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.*

Capítulo 2

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 desarollo 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 deseasemos realizar un análisis de demanda por curso, fecha y hora podria 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 3

Objetivos

3.1. 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.

3.2. 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 (análisis por curso y global).
- Determinar la distribución del número de alumnos asistentes a sesiones (análisis por curso y global).
- Determinar la relación entre el número de alumnos inscritos y el número de alumnos asistentes a sesiones (análisis por curso y global).
- Determinar la distribución del número de alumnos asistentes dado un número dado el número de alumnos asistentes (análisis por curso y global).

Capítulo 4

Trabajo realizado

4.1. Modelo matemático

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

Notación 4.1.1. Denotamos al conjunto de enteros en el intervalo $I \subset \mathbb{R}$ como $I_{\mathbb{Z}} := I \cap \mathbb{Z}$.

Definición 4.1.2 (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 4.1.3 (Cursos)**.** Definimos el conjunto de cursos como

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

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

Definición 4.1.4 (Tipo de sesión)**.** Definimos el conjunto de los tipos de sesiones como

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

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

Definición 4.1.5 (Sesión)**.** Definimos el conjunto de sesiones como

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

donde $\mathcal{D}_\mathcal{S} \subset \mathcal{D}$ y $\mathcal{H}_\mathcal{S} \subset \mathcal{H}$.

Definición 4.1.6 (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 4.1.7 (Número aleatorio de alumnos inscritos y asistentes)**.** Definimos el número de alumnos inscritos como la variable aleatoria $\mathcal{I} : \Omega \longrightarrow \mathbb{N}_0$ y el número de alumnos asistentes como la variable aleatoria $\mathcal{A} : \Omega \longrightarrow \mathbb{N}_0$.

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 8.0.9), pues resulta de una aplicación particular a una variable aleatoria.

Ejemplo 4.1.8 (Aproximación de la distribución de probabilidad)**.** Sea $Y : \Omega \longrightarrow \mathbb{R}$ una variable aleatoria y sea $X_1, X_2, X_3, \ldots$ 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

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

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

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

entonces por la ley fuerte de los grandes números se obtiene

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

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 $\mathbb{P}\left(Y \in A\right)$ 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

$$\mathbb{P}\left(Y \in A\right) \approx \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_A(X_i) \tag{4.1}$$

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

Ahora, con 4.1 estamos en condiciones de poder realizar las aproximaciones necesarias para el análisis exploratorio de los datos.

Observación 4.1.9. *Aproximamos las distribuciones de $\mathcal{I}$ y $\mathcal{A}$ mediante*

$$\mathbb{P}(\mathcal{I} = n) \approx \frac{\sum_{s\in\mathcal{S}} \mathbb{1}_{\{n\}}(I_s)}{\mathsf{n}(\mathcal{S})}, \quad \mathbb{P}(\mathcal{A} = n) \approx \frac{\sum_{s\in\mathcal{S}} \mathbb{1}_{\{n\}}(A_s)}{\mathsf{n}(\mathcal{S})} \tag{4.2}$$

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

Como $\mathsf{n}(\mathcal{S})$ 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 4.1.10 (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 4.2.

Notación 4.1.11. Denotamos $\hat{\mathbb{P}}_A(k) := \dfrac{\sum_{s\in\mathcal{S}} \mathbb{1}_{\{k\}}(a_s)}{\mathsf{n}(\mathcal{S})}$ para $k \in A$ y $a_s \in A$ una de las componentes de $s \in \mathcal{S}$.

Observación 4.1.12. *Aproximamos las distribuciones de $T_{\mathcal{S}}$, C, $D_{\mathcal{S}}$ y $H_{\mathcal{S}}$ mediante*

$$\mathbb{P}(T_{\mathcal{S}} = t) \approx \hat{\mathbb{P}}_{\mathcal{T}_{\mathcal{S}}}(t), \quad \mathbb{P}(C = c) \approx \hat{\mathbb{P}}_{\mathcal{C}}(c), \quad \mathbb{P}(D_{\mathcal{S}} = d) \approx \hat{\mathbb{P}}_{\mathcal{D}_{\mathcal{S}}}(d), \quad \mathbb{P}(H_{\mathcal{S}} = h) \approx \hat{\mathbb{P}}_{\mathcal{H}_{\mathcal{S}}}(h) \tag{4.3}$$

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

4.2. 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 4.2.1. *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 4.2.2. En motivación de la Observación 4.2.1 usamos la notación

$$\mathbb{P}_B(A) := \mathbb{P}(A \mid B) \tag{4.4}$$

para A y B eventos de un espacio de probabilidad.

Observación 4.2.3. *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*

$$\mathbb{P}(\mathcal{I} = n) = \sum_{c \in \mathcal{C}} \mathbb{P}(C = c)\, \mathbb{P}(\mathcal{I} = n \mid C = c) \tag{4.5}$$

y por lo tanto podemos obtener la aproximación de la distribución de $\mathcal{I}$ conociendo las aproximaciones de las distribuciones de C e 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 4.5 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 4.2.4 (Aproximación de la varianza)**.** Sea $Y : \Omega \longrightarrow \mathbb{R}$ una variable aleatoria y sea $X_1, X_2, X_3, \ldots$ 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

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

y si definimos

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

entonces por la ley fuerte de los grandes números se obtiene

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

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

$$\sigma^2 \approx \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 \tag{4.6}$$

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

Si bien podemos contentarnos con la expresión obtenida en 4.6, 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.

4.3. Métricas estadísticas

Definición 4.3.1 (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 4.3.2 (Contraste de la proporción (dos colas))**.** Sean las hipótesis:

$$H_0 : p = p_0, \quad H_1 : p \neq p_0,$$

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.

Definición 4.3.3 (Kurtosis)**.**

Definición 4.3.4 (Exceso de kurtosis)**.**

Definición 4.3.5 (Coeficiente de correlación (Pearson))**.**

Definición 4.3.6 (Matriz de correlación)**.**

4.4. *ETL* (Extracción-Transformación-Carga)

4.4.1. 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.

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 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.

4.4.2. 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**. 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*.

Observación 4.4.1. *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 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 4.4.1). 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.

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*.

4.4.3. 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*.

4.5. EDA (Análisis Exploratorio de los Datos)

En general, para realizar el análisis exploratorio de los datos 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.

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

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

Asimismo, se presentan los cursos dictados:

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

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

4.5.1. Tipo de sesión

Aquí la variable aleatoria analizada es $T_{\mathcal{S}}$. Esta presenta dos posibles valores: Taller o Tutoria, por lo que se trata de una variable categórica. Al existir dos posibles valores, se esperaría que la variable aleatoria $T_{\mathcal{S}}$ sea uniformemente distribuida, es decir,

$$\mathbb{P}\left(T_{\mathcal{S}} = \text{Taller}\right) = \mathbb{P}\left(T_{\mathcal{S}} = \text{Tutoria}\right) = \frac{1}{2} \tag{4.7}$$

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 $T_{\mathcal{S}}$.

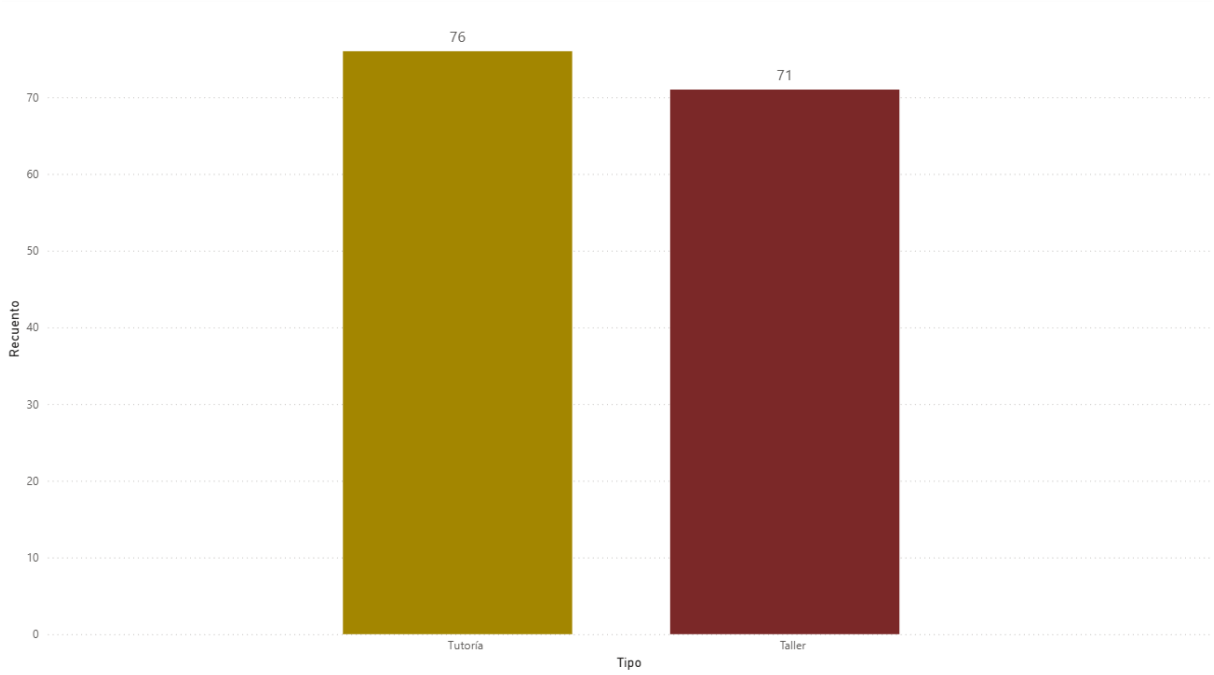

Figura 4.1: Gráfico de barras del tipo de sesión

Con ello formamos la tabla

t_s	$\hat{\mathbb{P}}_{\mathcal{T}_\mathcal{S}}(t_s)$
Taller	$\frac{71}{147}$
Tutoria	$\frac{76}{147}$

Tabla 4.1: Aproximación de la distribución de $T_\mathcal{S}$

Entonces, de acuerdo a lo planteado en 4.7 y usando las aproximaciones planteadas en 4.3 se puede comprobar que la aproximación obtenida de la distribución de $T_\mathcal{S}$ es correcta pues de acuerdo a la gráfica de la Figura 4.1.

$$\hat{\mathbb{P}}_{\mathcal{T}_\mathcal{S}}(\text{Taller}) = \frac{76}{147} \approx 0.517, \quad \hat{\mathbb{P}}_{\mathcal{T}_\mathcal{S}}(\text{Tutoria}) = \frac{71}{147} \approx 0.483 \tag{4.8}$$

Ahora, veamos qué sucede con la distribución de $T_\mathcal{S}$ para cada uno de los cursos, es decir, $T_\mathcal{S}$ dado C.

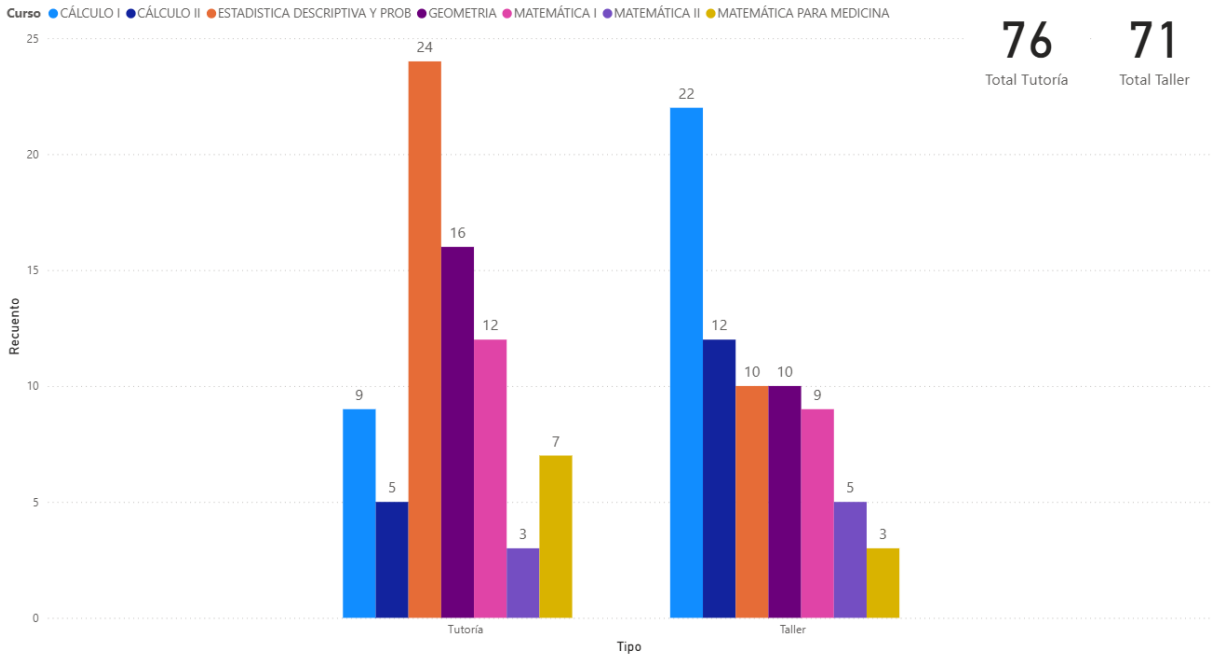

Figura 4.2: Gráfico de barras del tipo de sesión por curso

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

$\hat{\mathbb{P}}_{C=c}(T_{\mathcal{S}} = t_s)$		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 4.2: Aproximación de la distribución de $T_{\mathcal{S}}$ dado C

4.5.2. Cursos dictados

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 4.5. 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 el siguiente gráfico muestra un gráfico de barras obtenido a partir de *MeetingsCleanInfo* que resume la aproximacion obtenida de la distribución de la variable aleatoria C.

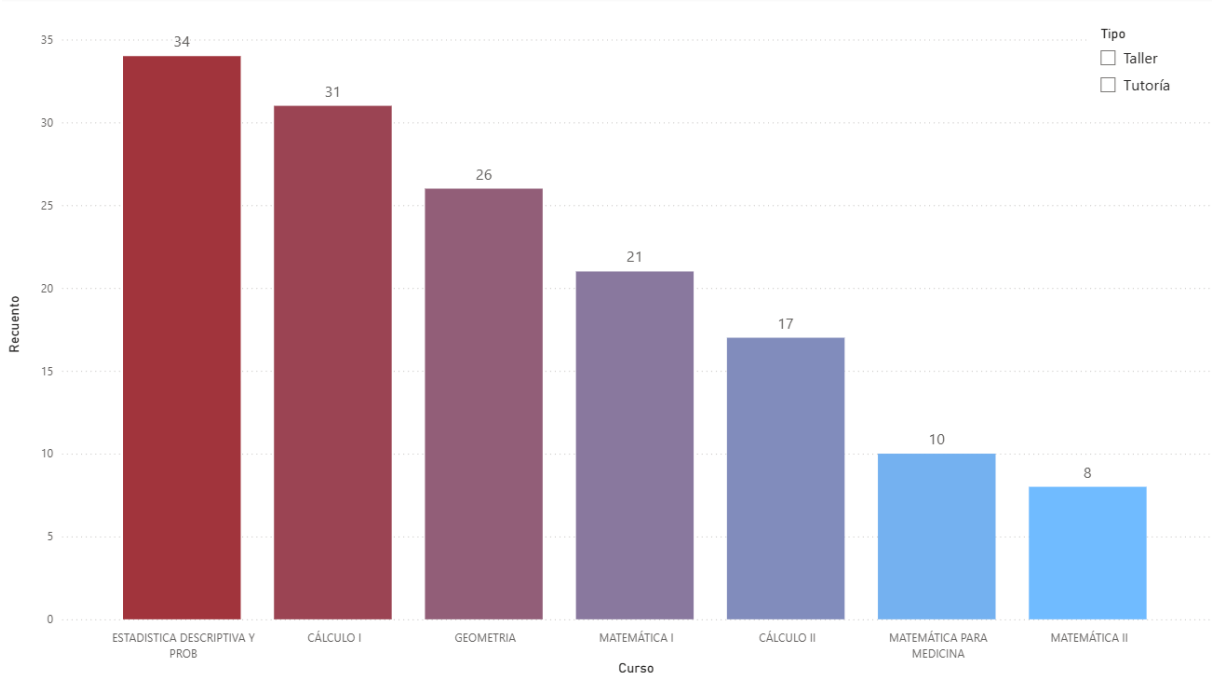

Figura 4.3: Histograma de los cursos dictados

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 4.3: Aproximación de la distribución de C

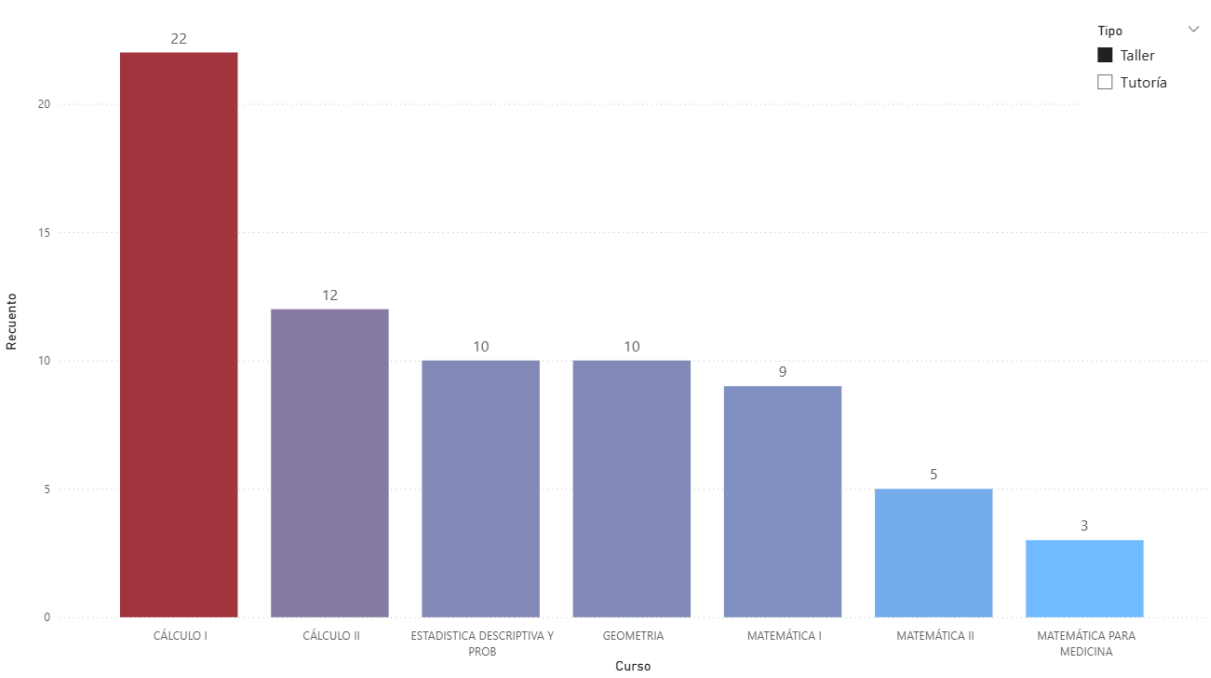

Figura 4.4: Histograma de los cursos dictados en talleres

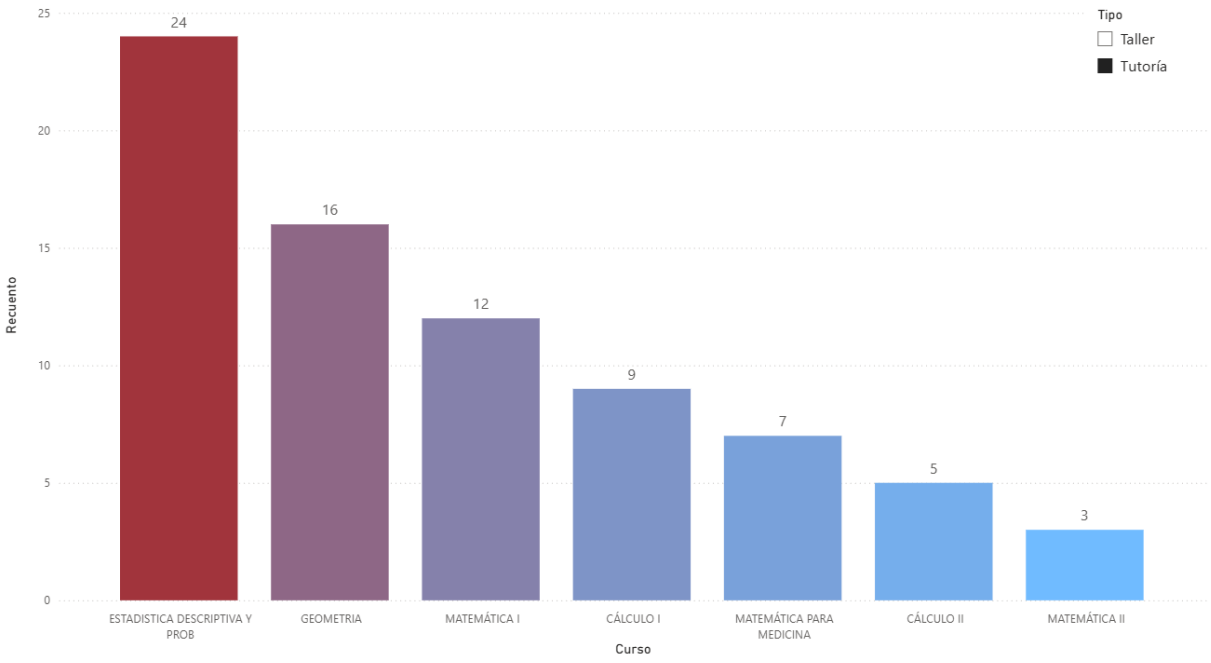

Figura 4.5: Histograma de los cursos dictados en tutorías

En base a esto último, mostramos la tabla de la aproximación de la distribución de C dado $T_{\mathcal{S}}$.

Se resalta el hecho de que $\mathsf{n}(\mathcal{S} \cap \{s \mid t_s = \text{Taller}\}) = 76$ y $\mathsf{n}(\mathcal{S} \cap \{s \mid t_s = \text{Tutoria}\}) = 71$ ya que estos valores aparecen en la parte superior de la Figura 4.1.

$\hat{\mathbb{P}}_{T_{\mathcal{S}}=t_s}(C=c)$		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 4.4: Aproximación de la distribución de C dado $T_{\mathcal{S}}$

4.5.3. Alumnos inscritos

Como sabemos, $\mathcal{I}$ representa la variable aleatoria número de alumnos inscritos en una sesión y mediante 4.2 se puede obtener su distribución aproximada, o también mediante 4.5.

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 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.

Incluso si los talleres solo permiten un máximo de 100 alumnos, 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 regristra 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.

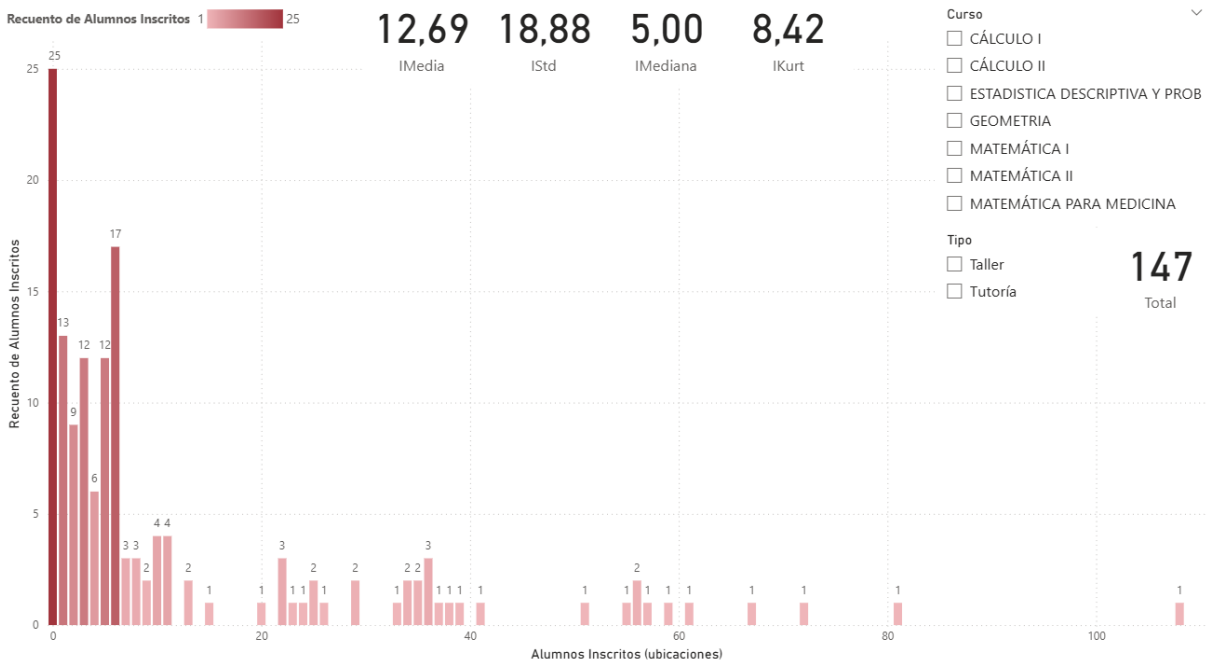

Figura 4.6: Histograma de inscritos

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.

4.5.4. Alumnos participantes

Como sabemos, $\mathcal{A}$ representa la variable aleatoria número de alumnos participantes o asistentes en una sesión y mediante 4.2 se puede obtener su distribución aproximada, o también mediante 4.5.

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 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.

En este caso aparentemente no se observan valores atípicos visibles (a diferencia de lo mostrado en la Figura 4.6).

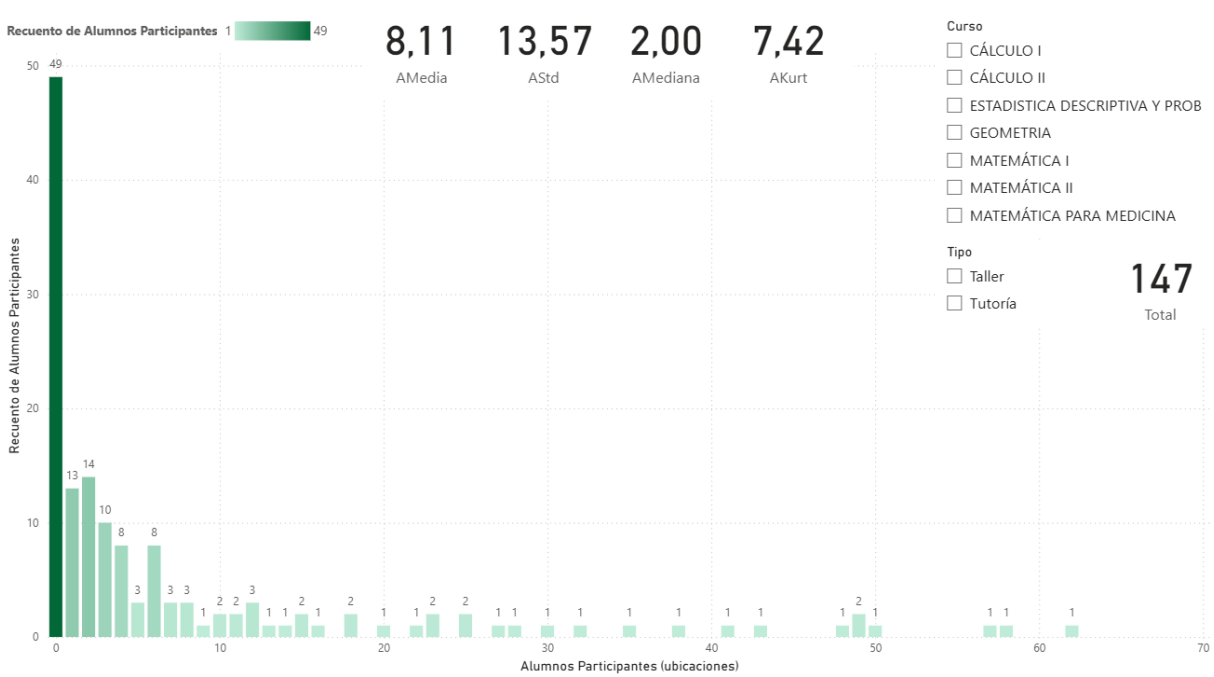

Figura 4.7: Histograma de participantes

Lo que sí es apreciable es que la tendencia en la distribución de A_s es bastante similar a la de I_s.

4.5.5. 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 4.8: 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 el siguiente gráfico 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*.

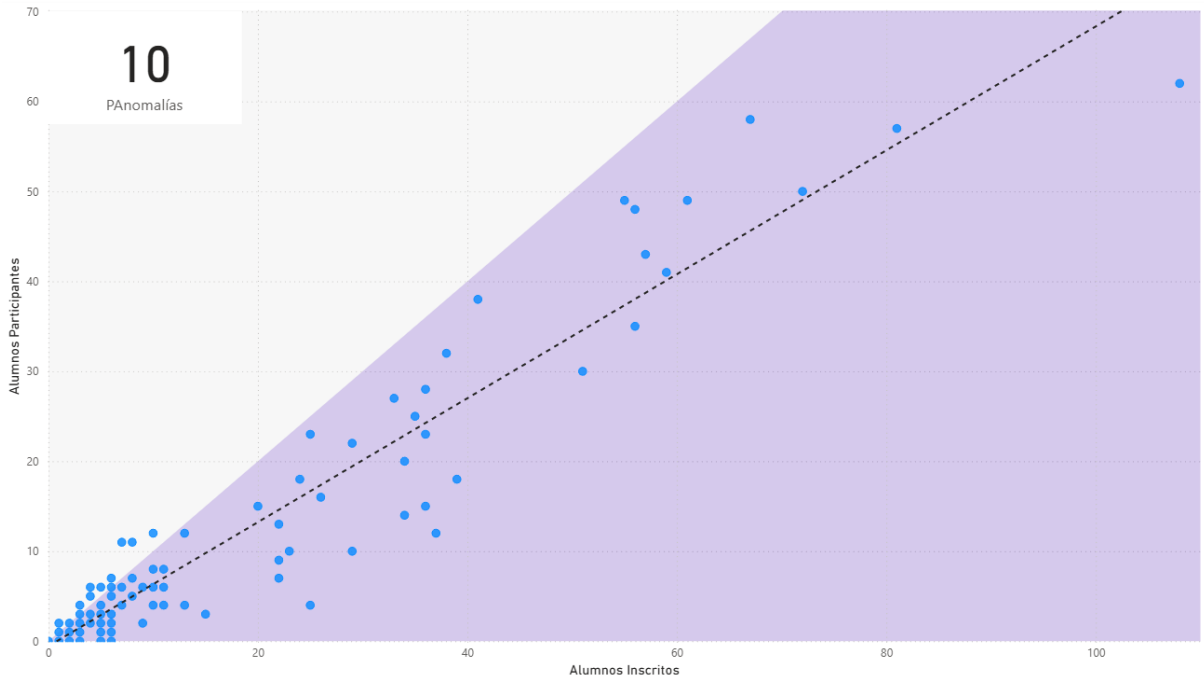

Figura 4.9: Dispersión de $(\mathcal{I}_s, \mathcal{A}_s)$ para cada $s \in \mathcal{S}$

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 I y A es bastante alta, podemos usar la regresión lineal para obtener una aproximación de la distribución de A dado I. En la Figura 4.9 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.

Así, la ecuación de la recta de regresión lineal obtenida se muestran en el siguiente gráfico.

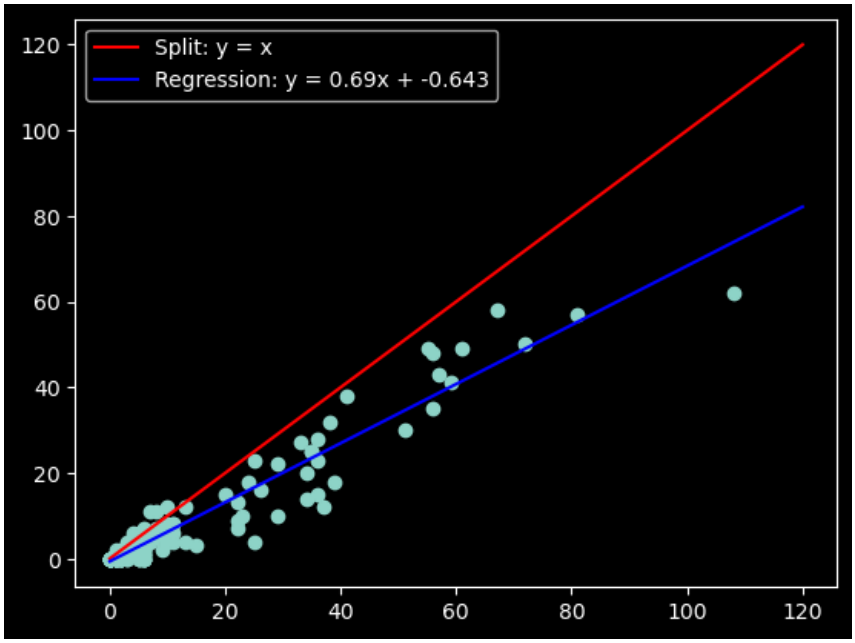

Figura 4.10: Regresión lineal de la variable aleatoria $\mathcal{I}, \mathcal{A}$

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 $A \approx 0.69I - 0.643 + \varepsilon_I$ con $\varepsilon_I \sim \mathcal{N}(0, \sigma_I^2)$ entonces

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

y deberíamos ser capaces de verificar que $A - 0.69I + 0.643 \sim \varepsilon_I \sim \mathcal{N}(0, \sigma_I^2)$ mediante el contraste de hipótesis.

4.5.6. Duración de sesiones

La columna **Duración (minutos)** 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

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

En el comportamiento global de esta variable podría no ser apreciable algún comportamiento en particular.

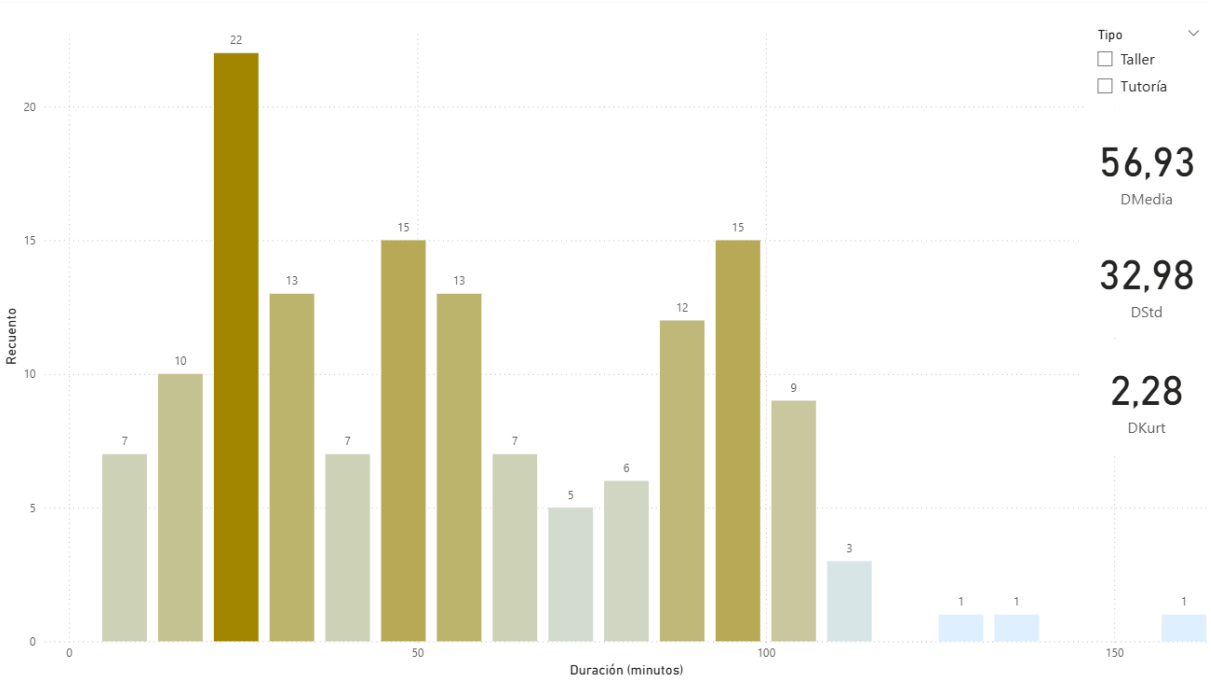

Figura 4.11: Histograma de la duración de las sesiones

Sin embargo, como ya se ha mencionado, según el tipo de sesión la duración máxima establecida de la misma es diferente. Es decir, 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.

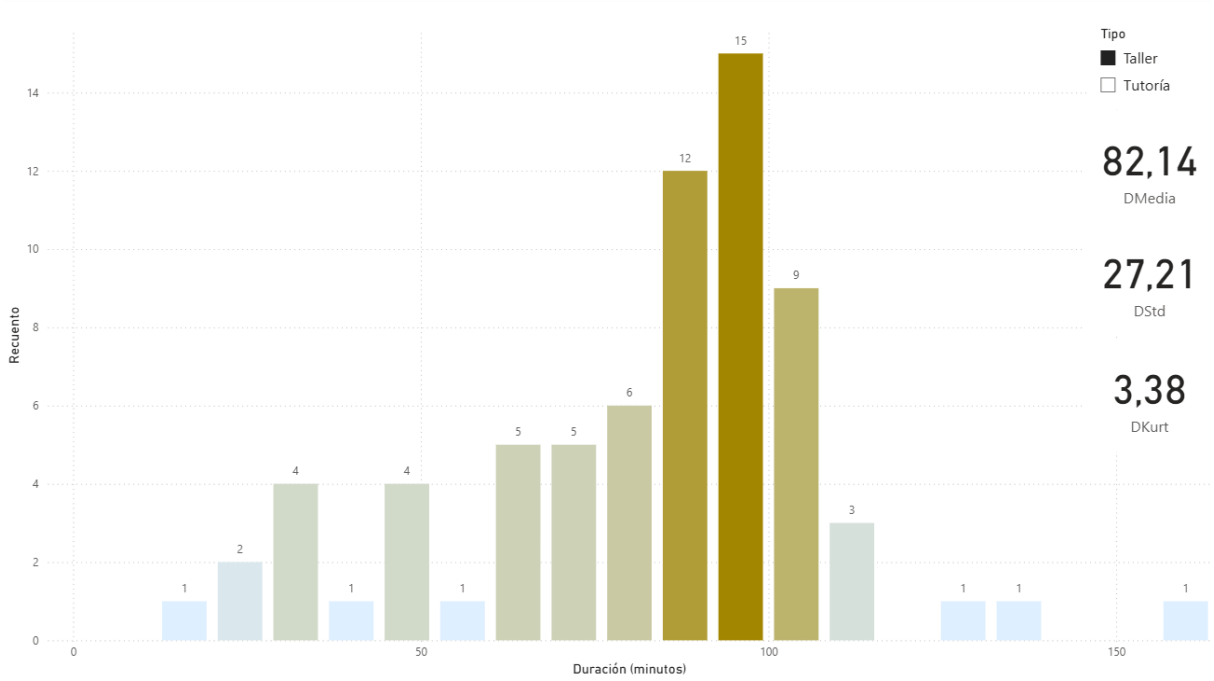

Figura 4.12: Histograma de la duración de las sesiones en talleres

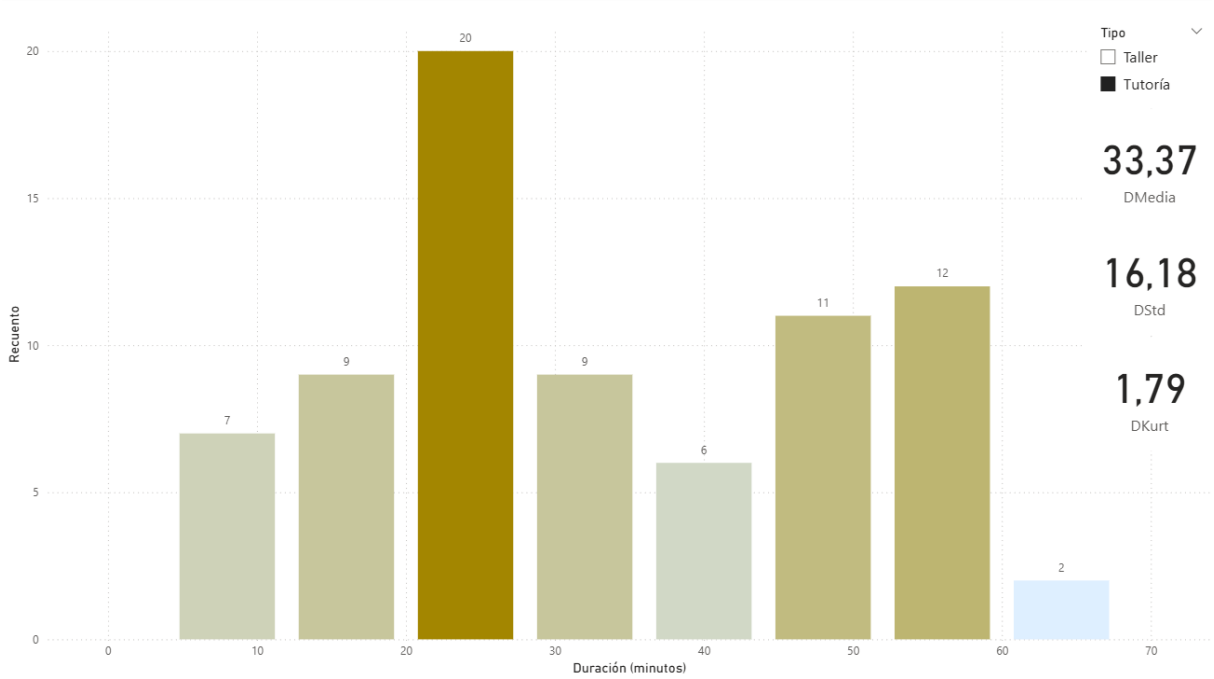

Figura 4.13: Histograma de la duración de las sesiones en tutorías

4.5.7. Puntualidad

La columna **Puntualidad (minutos)** 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

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

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 $[5; 10]$.

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.

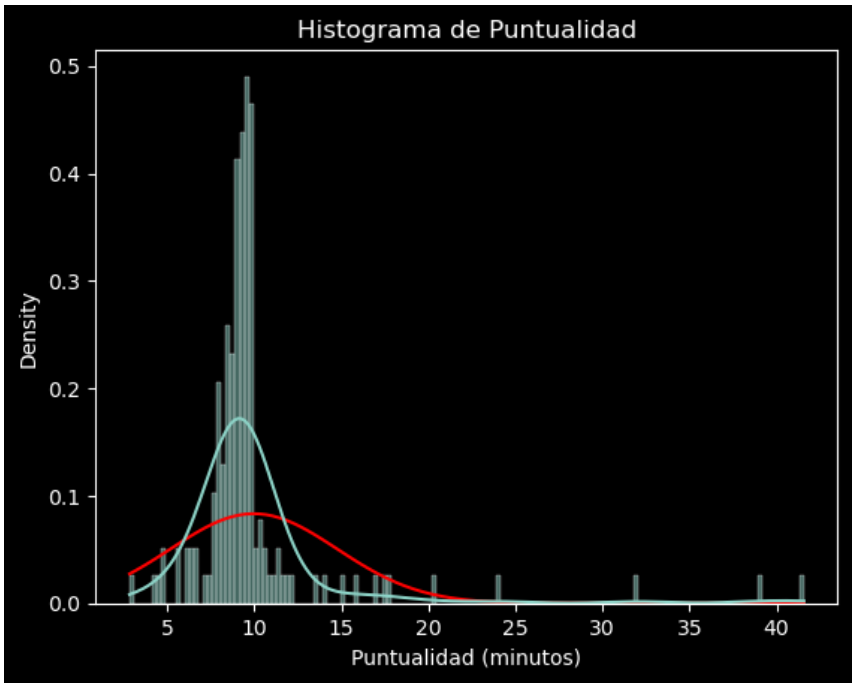

Figura 4.14: Histograma de la puntualidad

En efecto, los valores atípicos se aparecen y estos son los valores 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.

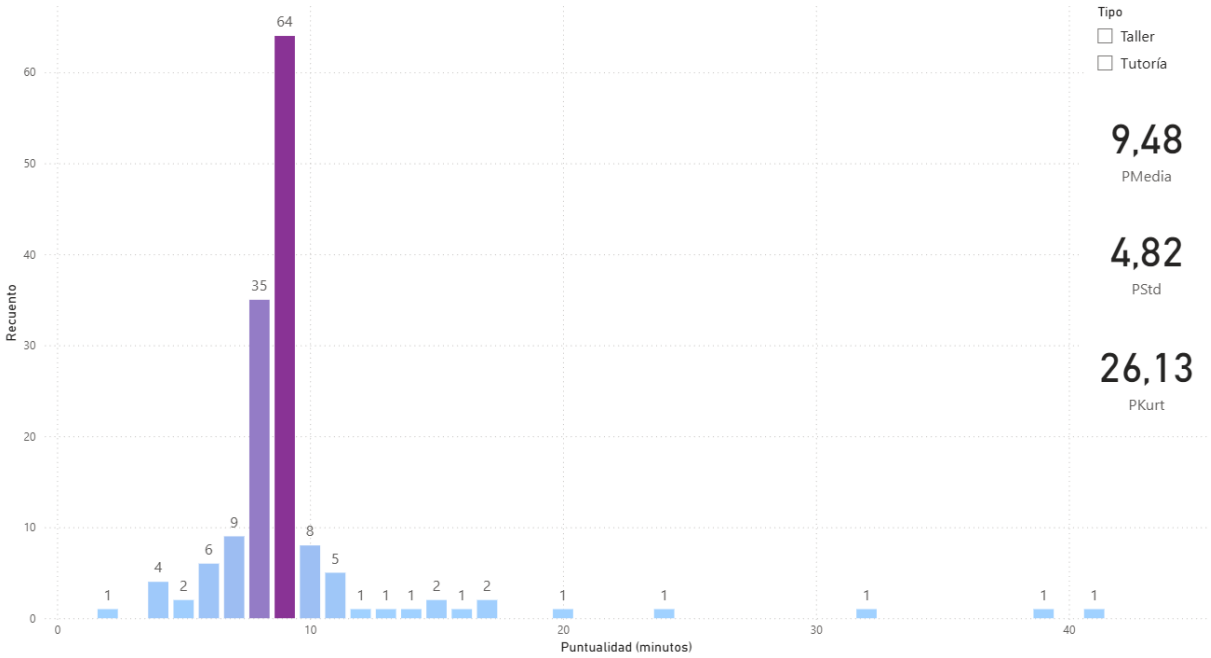

Figura 4.15: Histograma de la puntualidad

El comportamiento global es muy similar al observado al comportamiento en talleres y tutorías.

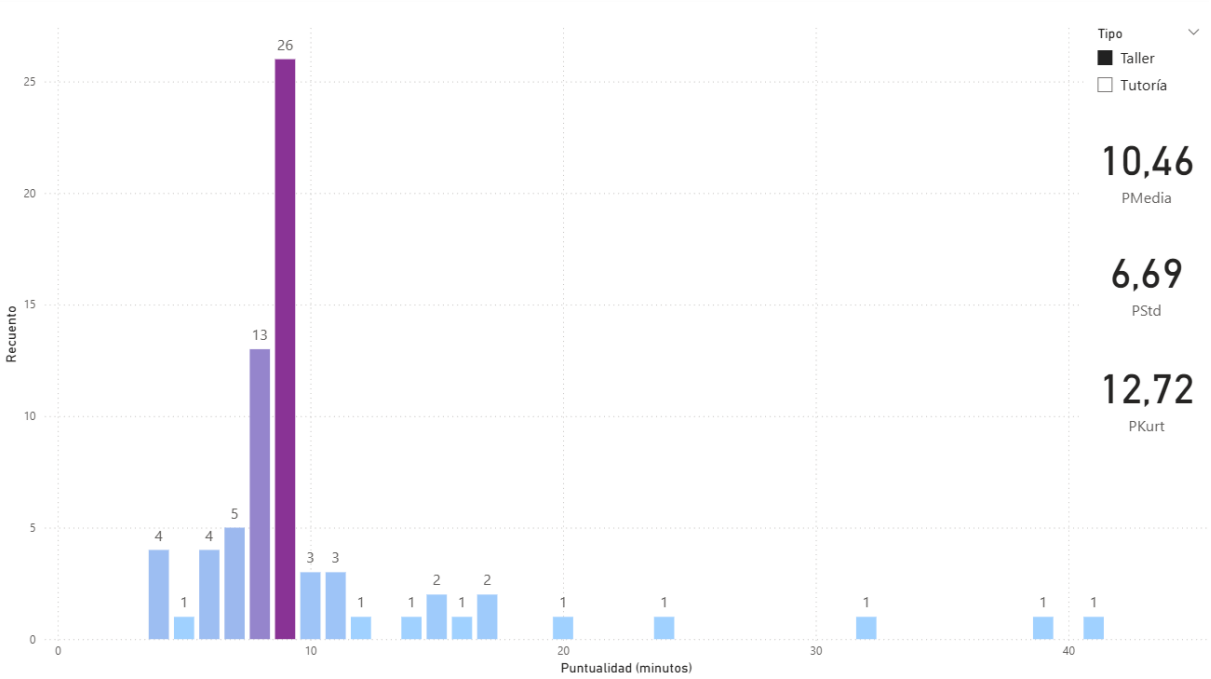

Figura 4.16: Histograma de la puntualidad

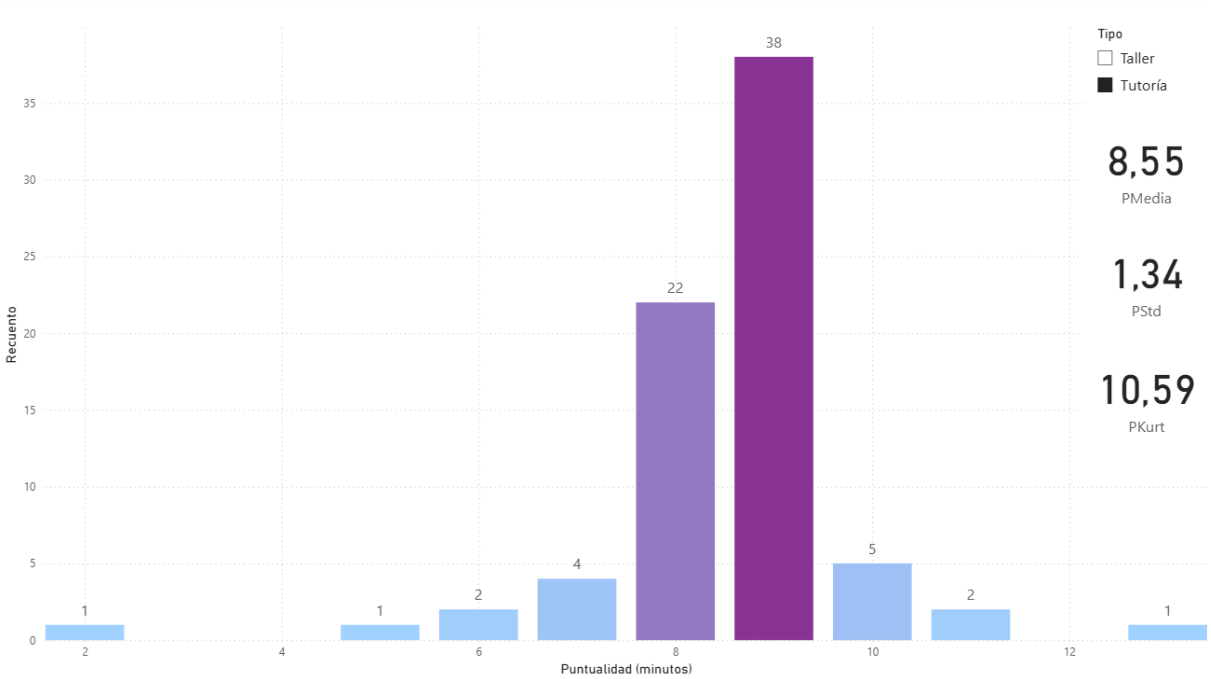

Figura 4.17: Histograma de la puntualidad

Esto tiene sentido, pues, debido a que la puntualidad es indistinta al tipo de sesión.

Capítulo 5

Resultados

Distribución del número de inscritos (Por curso y global: el número global es la suma de los números por curso (variables aleatorias))

5.1. Por curso

5.1.1. Matemática I

5.1.2. Matemática II

5.1.3. Cálculo I

5.1.4. Cálculo II

5.1.5. Estadística Descriptiva

5.1.6. Geometría

5.2. Por tipo

5.2.1. Talleres

5.2.2. Tutorías

5.3. Por fecha y hora

Capítulo 6

Discusión

Capítulo 7

Conclusiones

Capítulo 8

Anexos

Definición 8.0.1 (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

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

Definición 8.0.2 (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

$$\mathbb{P}(A \mid B) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \tag{8.1}$$

Definición 8.0.3 (Partición)**.** Decimos que los $A_1, A_2, \ldots, A_n \subset A$ forman una partición de A si los A_i son exhaustivos y mutuamente excluyentes.

Definición 8.0.4 (Partición de un espacio de probabilidad)**.** Decimos que los $A_1, A_2, \ldots, A_n \subset A$ forman una partición de un espacio de probabilidad $(\Omega, \mathcal{F}, \mathbb{P})$ si $A_i \in \mathcal{F}$, $i = 1, \ldots, n$ y forman una partición de Ω.

Teorema 8.0.5 (Bayes)**.** *Sea $A_1, A_2, \ldots, A_n$ una partición del espacio de probabilidad $(\Omega, \mathcal{F}, \mathbb{P})$ tales que $\mathbb{P}(A_i) > 0$, $i = 1, \ldots, n$. Sea B un evento arbitrario de $(\Omega, \mathcal{F}, \mathbb{P})$, entonces*

$$\mathbb{P}(A_i \mid B) = \frac{\mathbb{P}(B \mid A_i)\,\mathbb{P}(A_i)}{\mathbb{P}(B)} \tag{8.2}$$

Teorema 8.0.6 (Probabilidad total)**.** *Bajo las mismas condiciones del Teorema 8.0.5,*

$$\mathbb{P}(B) = \sum_{i=1}^{n} \mathbb{P}(B \mid A_i)\mathbb{P}(A_i) \tag{8.3}$$

Además se obtiene la ***fórmula de Bayes***

$$\mathbb{P}(A_i \mid B) = \frac{\mathbb{P}(B \mid A_i)\,\mathbb{P}(A_i)}{\sum_{i=1}^{n} \mathbb{P}(B \mid A_i)\mathbb{P}(A_i)} \tag{8.4}$$

Definición 8.0.7 (Distribución normal)**.** Una variable aleatoria X es normalmente distribuida con media μ y varianza σ^2 si su densidad es

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{8.5}$$

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

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

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

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

Teorema 8.0.9 (Ley fuerte de los grandes números)**.** *Sea $X_1, X_2, X_3, \ldots$ 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*

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

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

Definición 8.0.10 (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 8.0.11 (Sesgo de un estimador)**.** Sea $\hat{\theta}$ un estimador de un parámetro θ. Entonces el sesgo o *bias* de $\hat{\theta}$ se define como

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

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

Definición 8.0.12 (Coeficiente de correlación)**.** .

Bibliografía

[AL16] Sylvain Arlot y Matthieu Lerasle. «Choice of V for V-Fold Cross-Validation in Least-Squares Density Estimation». En: *Journal of Machine Learning Research* 17.208 (2016), págs. 1-50. URL: http://jmlr.org/papers/v17/14-296.html.

[BKM14] Adil Bagirov, Napsu Karmitsa y Marko M. Mkel. *Introduction to Nonsmooth Optimization: Theory, Practice and Software*. Springer Publishing Company, Incorporated, 2014.

[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.

[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.

[Unk] Unknown. *Subgradient of Convex Function*. URL: https://www.math.cuhk.edu.hk/course_builder/1920/math4230/Note8.pdf.

[UTP] UTP. *UTP Reservas*. URL: https://reservarecursos.utp.edu.pe/ref-acad.