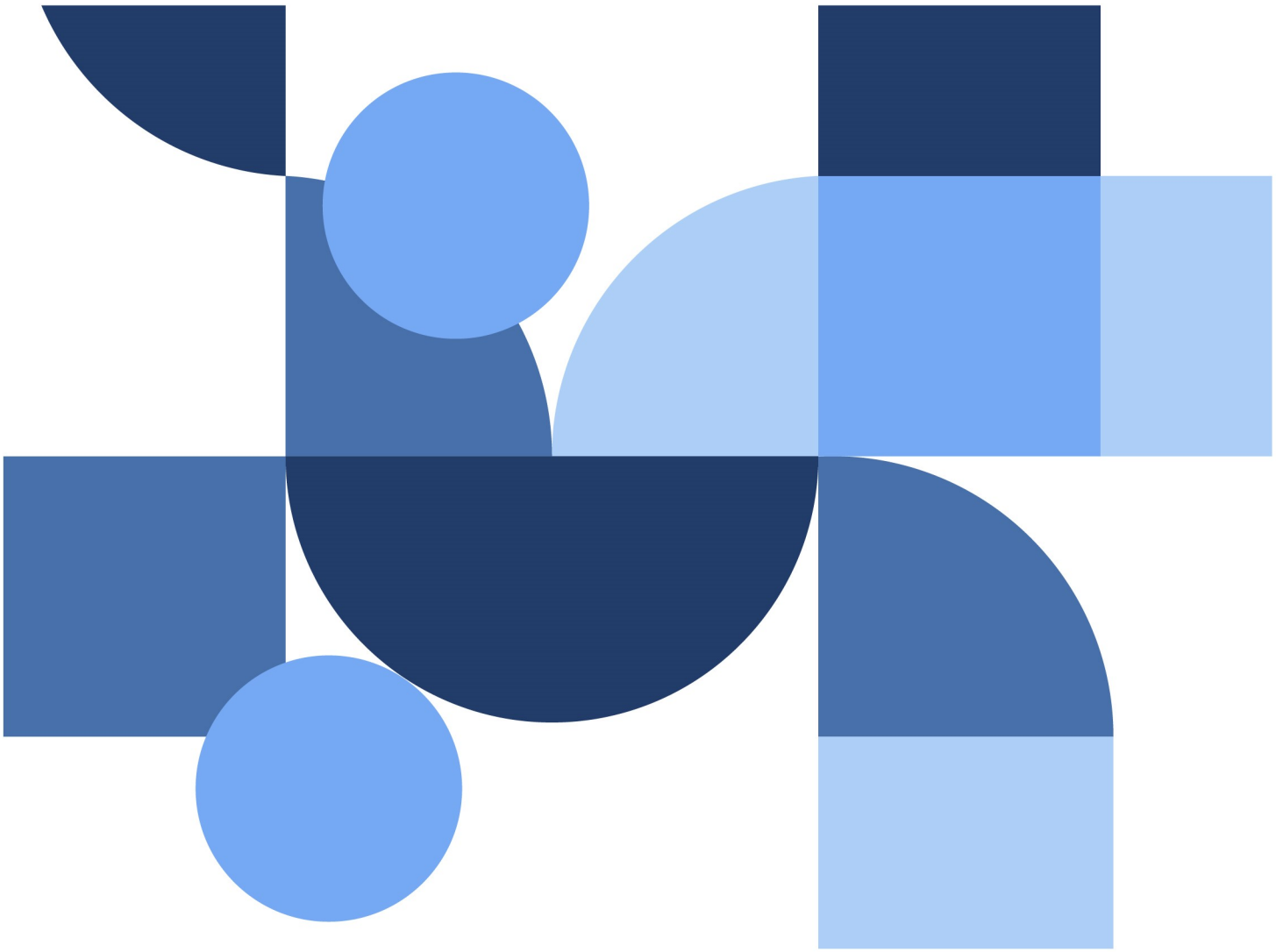
Dirección de Docencia
Universidad de Concepción



Enseñanza y aprendizaje de la estadística descriptiva e inferencial

Autores:
Luisa Rivas Calabrán
Eugenio Chandía Muñoz

Enseñanza y Aprendizaje de la Estadística Descriptiva e Inferencial

LUISA RIVAS CALABRÁN
EUGENIO CHANDÍA MUÑOZ

Registro de Propiedad Intelectual N° 2026-P-4710
ISBN: 978-956-9280-62-7

Agradecimientos

La elaboración de este libro fue posible gracias al apoyo del proyecto Escribe.doc, impulsado por la Dirección de Docencia de la Universidad de Concepción. Esta iniciativa busca generar recursos didácticos digitales que acompañen la labor del profesorado y favorezcan el aprendizaje de los estudiantes. En este sentido, lo que aquí compartimos nace de nuestra propia práctica docente y del deseo de poner a disposición de la comunidad educativa materiales útiles, significativos y accesibles para enriquecer la enseñanza y el aprendizaje de la Estadística y la Probabilidad.

Este texto tiene como principal objetivo fortalecer la formación de los estudiantes de Pedagogía en Matemáticas en los contenidos del eje de Estadística y Probabilidad, ofreciendo una herramienta didáctica y disciplinar que les permita reflexionar sobre la enseñanza de estos temas y mejorar su capacidad para impartirlos de manera efectiva en la educación media, conforme a los planes y programas del Ministerio de Educación.

Nuestro agradecimiento se dirige, en primer lugar, a la Universidad de Concepción por fomentar iniciativas que buscan dar respuesta a necesidades formativas específicas, en áreas que han sido históricamente postergadas en la educación chilena. En particular, reconocemos la relevancia que este proyecto adquiere frente a los resultados de la Evaluación Nacional Diagnóstica de la Formación Inicial Docente (End-FID), que han puesto de manifiesto las debilidades en el eje de Datos y Azar, así como la escasez de orientaciones y materiales que permitan planificar y ejecutar clases centradas en el pensamiento estadístico.

Agradecemos también a las y los estudiantes de Pedagogía en Matemáticas, cuyas experiencias, autopercepciones y desafíos en torno a la estadística escolar motivaron la creación de este recurso. Ellos han sido, en última instancia, la razón fundamental para diseñar un material que contribuya a superar las carencias detectadas en su formación inicial y los prepare mejor para su futuro rol docente.

De manera especial, expresamos nuestro reconocimiento a las y los colegas que con su retroalimentación académica y pedagógica enriquecieron la propuesta, y a quienes han impulsado la investigación en torno a la enseñanza de la probabilidad y la estadística, cuyos hallazgos sirvieron de base para orientar este esfuerzo.

Finalmente, extendemos nuestro agradecimiento a todas las personas e instancias que hicieron posible la concreción de este libro, cuyo propósito es aportar al fortalecimiento del pensamiento estadístico de los futuros docentes y, con ello, contribuir a mejorar la calidad de la enseñanza de la Estadística y la Probabilidad en las aulas chilenas.

Luisa Rivas Calabrán
Facultad de Ciencias Físicas y Matemáticas
Departamento de Estadística
Universidad de Concepción

Eugenio Chandía Muñoz
Facultad de Educación
Departamento de Currículum e Instrucción
Universidad de Concepción

Prólogo

La enseñanza de la Estadística y la Probabilidad en Chile constituye un desafío relativamente reciente dentro de la educación matemática escolar. Aunque en 2009 se implementó un ajuste curricular que integró este eje desde la Educación Básica hasta la Enseñanza Media, con el propósito de dotar a los estudiantes de herramientas para comprender la información en diversos contextos y evaluar críticamente estudios basados en datos, la realidad ha demostrado que este camino aún se encuentra en construcción. La incorporación, en 2020, del electivo de Probabilidades y Estadística Descriptiva e Inferencial en Tercero y Cuarto Medio fue un avance significativo, pero no suficiente para revertir las dificultades observadas en su enseñanza y aprendizaje.

Uno de los indicadores más claros de esta situación lo constituye la Evaluación Nacional Diagnóstica de la Formación Inicial Docente (End-FID), que ha mostrado consistentemente un bajo desempeño en el eje de Datos y Azar. En los años 2017, 2019 y 2021, los porcentajes de respuestas correctas en este eje alcanzaron apenas un 37,3 %, 47,4 % y 44,8 % respectivamente, revelando que los futuros profesores de matemáticas presentan falencias importantes en su formación estadística inicial. Estas cifras no son un simple dato numérico: reflejan una carencia que se proyecta hacia la enseñanza escolar, ya que los profesores en formación tienden a enfrentar sus cursos de estadística con un bagaje de conocimientos limitado, percepciones de inseguridad en torno a la disciplina y escasas estrategias didácticas para enseñar sus contenidos.

La investigación en educación matemática confirma este diagnóstico. Se ha demostrado que los futuros docentes, si bien manifiestan una actitud generalmente positiva hacia la probabilidad y la estadística, no siempre cuentan con una preparación adecuada para enseñarlas de manera significativa (Ruz et al., 2020). Otros estudios muestran que sus habilidades didácticas son insuficientes (Vásquez & Alsina, 2014) y que, en la práctica escolar, las actividades relacionadas con probabilidad y estadística suelen ser mecánicas, centradas en procedimientos, con poca atención a la comprensión conceptual y a la formación de un verdadero pensamiento estadístico (Vásquez & Alsina, 2021).

Este panorama evidencia una necesidad urgente: ofrecer a los futuros docentes de matemáticas oportunidades de formación que integren los contenidos disciplinares con su enseñanza en la escuela, desde una perspectiva reflexiva y aplicada. No basta con que los estudiantes de pedagogía conozcan fórmulas o procedimientos; es imprescindible que comprendan cómo esos conceptos se construyen, cómo se aplican en la vida cotidiana, y cómo pueden ser enseñados en contextos escolares de manera que promuevan el razonamiento y la interpretación crítica de datos.

Este libro surge precisamente de esa necesidad. Su propósito central es fortalecer la formación de los estudiantes de Pedagogía en Matemáticas en el eje de Estadística y Probabilidad, mediante un recurso que combina el rigor disciplinar con orientaciones didácticas. Lo que aquí se presenta no es un manual más de fórmulas o ejercicios mecánicos: se trata de una propuesta construida a partir de problemas abiertos y asimétricos que reflejan situaciones auténticas, en línea con los contenidos y objetivos de aprendizaje establecidos por el Ministerio de Educación en la última actualización de planes y programas.

Los problemas incluidos en este texto poseen una dificultad creciente y están diseñados para estimular la reflexión, el debate y la consolidación de aprendizajes. Cada capítulo propone un desafío inicial, seguido de elementos que apoyan la comprensión del estudiante: la planificación del problema (con su identificación curricular, enunciado, simplificación, extensión y posibles soluciones), el desarrollo de los contenidos disciplinares que lo sustentan, y finalmente, las orientaciones didácticas que guían su enseñanza. Este esquema permite a los futuros docentes experimentar el doble rol que caracterizará su ejercicio profesional: por un lado, resolver problemas como estudiantes de estadística; por otro, analizar cómo esos mismos problemas pueden enseñarse en el aula para promover aprendizajes significativos.

La estructura del libro está pensada para acompañar progresivamente este proceso formativo. En el capítulo

inicial se explican los lineamientos generales y la lógica de organización de los problemas. Los capítulos 2 al 8 constituyen el núcleo del texto, presentando problemas organizados según el esquema descrito. Adicionalemente, cada uno de estos capítulos finaliza con orientaciones que recogen las recomendaciones generales para la enseñanza de la estadística y la probabilidad en Tercero y Cuarto medio, incluyendo los errores más frecuentes observados en la resolución de los problemas propuestos. Finalmente, se presentan las referencias que respaldan teórica y metodológicamente esta propuesta. Y por último se pone a disposición del lector los problemas desarrollados en un formato más amigable para su implementación.

Al adoptar esta estructura, el libro busca subsanar una de las falencias más recurrentes en la formación inicial: la desconexión entre los contenidos matemáticos y su enseñanza. Aquí, cada problema funciona como un puente que vincula el conocimiento disciplinar con la reflexión didáctica, invitando a los futuros docentes a transitar del "cómo resolver" al "cómo enseñar".

Este texto está dirigido principalmente a estudiantes de Pedagogía en Matemáticas, pero también puede ser de utilidad para docentes en ejercicio que deseen fortalecer su práctica en el eje de Estadística y Probabilidad, así como para académicos y formadores de profesores que busquen materiales innovadores para acompañar sus cursos.

Invitamos a las y los lectores a recorrer estas páginas con espíritu crítico y reflexivo, explorando los problemas, analizando sus soluciones y, sobre todo, preguntándose cómo cada uno de ellos puede transformarse en una oportunidad de aprendizaje en el aula escolar. Nuestra aspiración es que este libro contribuya no solo a mejorar la formación inicial en estadística, sino también a consolidar un enfoque de enseñanza que forme estudiantes capaces de comprender el mundo a través de los datos, tomar decisiones informadas y participar activamente en una sociedad cada vez más atravesada por la información.

Confiamos en que este esfuerzo, nacido desde la práctica docente y respaldado por el proyecto Escribe.doc de la Universidad de Concepción, se convierta en un aporte significativo para la formación de los futuros profesores de matemáticas y, en consecuencia, para el fortalecimiento de la educación estadística en Chile.

Índice

1. Preliminares

1.1. Estructura de cada problema

Cada planificación del problema se organiza según la siguiente secuencia, la cual se ha diseñado para articular los objetivos curriculares, el contenido disciplinar y las oportunidades didácticas que favorecen el desarrollo del razonamiento estadístico y probabilístico:

1. **Identificación curricular y de contenido del problema.**

   Esta sección sintetiza tres componentes en la Tabla 1: (i) el *Objetivo de Aprendizaje (OA) del Currículo Escolar*; (ii) el *Objetivo de Aprendizaje de la Actividad* (definido por los autores); (iii) los *Contenidos de Estadística y Probabilidad* requeridos para resolver la actividad.

   Tabla 1: Mapa entre OA curriculares, OA de la actividad y contenidos de Estadística y Probabilidad

OA del Currículo Escolar	OA de la Actividad	Contenidos de Estadística y Probabilidad

   Para este libro, los OA curriculares provienen de las *Bases Curriculares* correspondientes a la Formación Diferenciada Humanista–Científico de Tercero y Cuarto Medio del Ministerio de Educación de Chile, específicamente de la asignatura *Probabilidades y Estadística Descriptiva e Inferencial*. El *OA de la Actividad* es propuesto por los autores, y los *Contenidos* detallan el conocimiento estadístico y probabilístico necesario para su implementación.

2. **Problema.** Se presenta una situación auténtica, acompañada de representaciones (gráficos, tablas o contextos reales) que demanda el análisis, interpretación y justificación de conclusiones. El problema busca vincular el contenido disciplinar con fenómenos significativos para los estudiantes, promoviendo así el uso de sus conocimientos previos y de su capital cultural.

3. **Una solución del problema.** Se entrega un conjunto de fundamentos y explicaciones que constituyen una propuesta de resolución. Esta sección orienta al profesorado respecto de los caminos plausibles que podrían recorrer los estudiantes.

4. **Una simplificación.** Se formulan interrogantes alternativas o ajustes en el nivel de dificultad del problema, con el fin de atender a la diversidad de desempeños de los estudiantes y facilitar procesos de acceso gradual al conocimiento.

5. **Consolidación del conocimiento o habilidades pretendidas.** Se proponen preguntas que permiten verificar si los estudiantes comprenden los conceptos y razonamientos involucrados, promoviendo la metacognición y el tránsito desde lo procedimental a lo conceptual.

6. **Una extensión.** Se incorpora información adicional (tablas, gráficos o datos complementarios) que amplía la problemática inicial y favorece la integración de nuevos elementos de análisis, estimulando la transferencia y la capacidad de inferencia.

7. **Plenaria.** Finalmente, se sugieren criterios para la selección de respuestas y estrategias de discusión colectiva. Esta instancia busca promover la socialización de diferentes enfoques, la validación de argumentos y la construcción de inferencias informales a partir de la evidencia.

En síntesis, cada problema constituye una *unidad didáctica estructurada* que vincula los objetivos curriculares con la práctica de aula. La secuencia de identificación, problema, solución, simplificación, consolidación, extensión y plenaria permite a los formadores de profesores y a los propios estudiantes transitar desde el reconocimiento de los contenidos estadísticos y probabilísticos hasta la construcción de inferencias y generalizaciones. Con ello, se favorece no solo el aprendizaje disciplinar, sino también el desarrollo del razonamiento estadístico, la interpretación crítica de datos y la toma de decisiones fundamentadas.

1.2. Sugerencias para la implementación de las actividades

El modelo de implementación utilizado en el proyecto que dio origen a este libro fue la **indagación colaborativa**. La colaboración en la formación docente actúa como un catalizador del desarrollo profesional al desplazar el foco desde el aprendizaje individual hacia el aprendizaje *con y a través de otros*, abriendo espacios para la indagación sobre la práctica, el contraste de perspectivas y la toma de decisiones basada en evidencia. Entendida como *postura profesional y ética*, la indagación colaborativa habilita juicios pedagógicos bien fundados y relaciones de confianza con los aprendizajes del estudiantado (Johnson Lachuk et al., 2019).

La colaboración efectiva, sin embargo, no ocurre por defecto: requiere un *diseño pedagógico intencionado* que desarrolle habilidades de trabajo en equipo (p. ej., entrenamiento explícito, facilitación en tiempo real, instancias de *check-in* y *check-out*, y un *lenguaje compartido* para hablar de la colaboración). La evidencia muestra que estas intervenciones mejoran los aprendizajes percibidos y la calidad de los procesos grupales; en cambio, formar grupos sin andamiaje reduce la probabilidad de aprendizaje productivo (Sjølie et al., 2021).

En Educación Matemática, la colaboración se articula con el *conocimiento para enseñar* (MKT/PCK). Desde Shulman (1986) y la formalización de MKT por Ball et al. (2008), sabemos que enseñar matemáticas exige, además del dominio del contenido, la capacidad de transformarlo didácticamente: interpretar el pensamiento estudiantil, seleccionar representaciones y tareas, y justificar decisiones instruccionales. Los entornos colaborativos hacen visible ese conocimiento en uso porque exigen explicar, argumentar y rediseñar propuestas frente a pares (Appova & Taylor, 2020).

1.3. Énfasis en estadística y probabilidad

En **estadística y probabilidad**, la indagación colaborativa es especialmente fértil por, al menos, tres razones:

1. **Pensamiento estadístico y razonamiento con datos.** El marco de Wild y Pfannkuch (1999a)—problematizar, planificar, analizar y concluir—exige argumentar con evidencia, escrutar supuestos y razonar sobre *variabilidad* e *incertidumbre*. Estas prácticas se potencian cuando las y los docentes en formación contrastan interpretaciones y negocian conclusiones en equipos que examinan datos reales o simulados.

2. **Conocimiento profesional específico (SKT/MKT estadístico).** Se requieren constructos análogos al MKT, como el *Statistical Knowledge for Teaching* (SKT), que integren conocimiento del contenido estadístico, del estudiantado y de la enseñanza. Marcos influyentes describen cómo este conocimiento se activa para interpretar respuestas, seleccionar representaciones (tablas, diagramas, simulaciones) y anticipar errores frecuentes (p. ej., equiprobabilidad mal inferida) (Burgess, 2009; Groth, 2007). La indagación colaborativa favorece que el SKT emerja y se refine mediante el diálogo entre pares.

3. **Diseño de tareas y progresiones de ideas.** La literatura sobre *statistical reasoning* recomienda secuencias que integren contenido, pedagogía y evaluación, con tareas que exijan justificar con evidencia, discutir la variabilidad y modelar el azar; trabajar estas tareas en comunidades colaborativas mejora la anticipación de trayectorias de aprendizaje y el ajuste de apoyos instruccionales (Garfield & Ben-Zvi, 2008).

En síntesis, una *infraestructura colaborativa intencionada* (facilitación, acuerdos, lenguaje compartido) y marcos de conocimiento profesional (MKT/SKT) permiten que las y los futuros docentes *aprendan a enseñar* estadística y probabilidad con base en evidencia, currículo y análisis del pensamiento estudiantil, en coherencia con avances del campo y experiencias efectivas en cursos de contenido (Appova & Taylor, 2020; Ball et al., 2008; Johnson Lachuk et al., 2019; Sjølie et al., 2021).

1.4. Modelo de gestión de la clase

Se proponen cuatro momentos (Ver Figura 1):

1. **Normas y conformación de equipos.** Presentar criterios de trabajo colaborativo y conformar equipos (idealmente de 3–4 integrantes) por asignación aleatoria o criterios pedagógicos. Explicitar instancias de socialización, externalización, combinación e internalización de ideas. El/la docente alterna aproximaciones *pasivas*

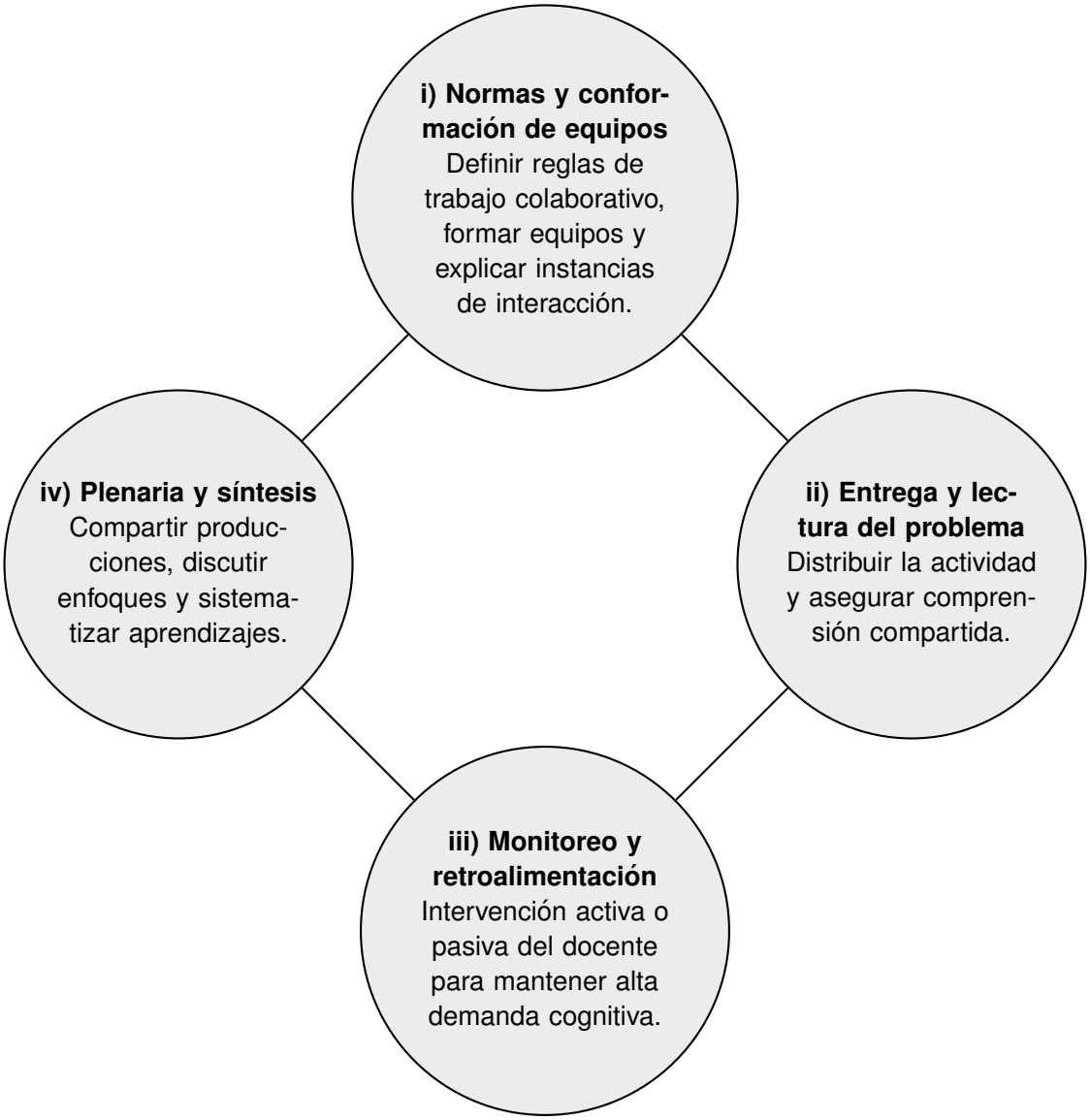


Figura 1: Modelo de gestión de la clase en cuatro momentos (estructura circular).

(observación y registro) y *activas* (intervención breve y focalizada cuando el equipo lo requiera o quede bloqueado).

2. **Entrega y lectura del problema.** Entregar la actividad (una por persona o una por equipo, si se desea forzar la interdependencia positiva). Asegurar una lectura compartida de consignas y una primera aproximación a ideas y representaciones.

3. **Monitoreo y retroalimentación durante el trabajo.** Definir cuándo y cómo el profesorado interviene (p. ej., bajo demanda del equipo). Mantener una *exigencia cognitiva alta* mediante preguntas que promuevan explicación, justificación y conexión entre representaciones. Si el equipo avanza con rapidez, ofrecer extensiones; si se estanca, proponer andamios o simplificaciones planificadas en el diseño de la actividad.

4. **Plenaria y síntesis.** Seleccionar estratégicamente producciones de equipos para la discusión pública (criterios: diversidad de enfoques, errores productivos, representaciones contrastantes). Formular un objetivo explícito para la plenaria (p. ej., comparar modelos, discutir supuestos, estimar y justificar) y cerrar con una *sistematización breve* que recupere conceptos, procedimientos y argumentos.

2. Problema 1

2.1. Planificación del Problema

a) **Identificación curricular y de contenido del problema**

Objetivo de Aprendizaje del Currículo Escolar	Objetivo de Aprendizaje de la Actividad	Contenidos Estadísticos y de Probabilidad
OA 1. Argumentar y comunicar decisiones a partir del análisis crítico de información presente en histogramas, polígonos de frecuencia, frecuencia acumulada, diagramas de cajón y nube de puntos, incluyendo el uso de herramientas digitales.	Analizar críticamente información presente en un gráfico de línea, identificando los elementos estructurales que permiten realizar extrapolaciones e interpolaciones.	■ Tipos de variable. ■ Gráfico de líneas.

b) **Problema** ¿Una carrera técnica o una profesional?

Observa el gráfico y explica por qué se producen las variaciones de búsqueda en el rango de tiempo.

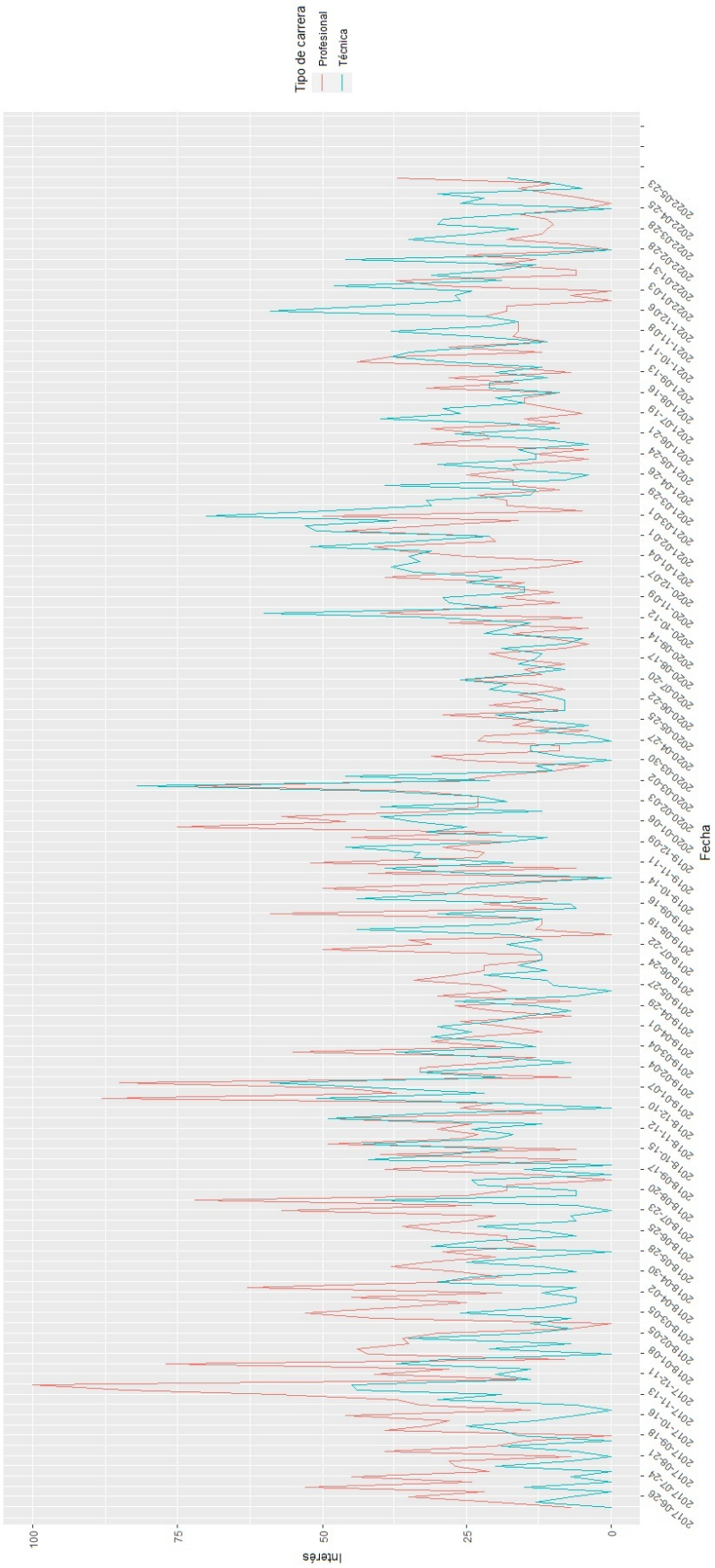


Figura 2: Búsqueda carreras técnicas y profesionales

c) **Una solución del problema**

Los fundamentos que pueden explicar la variación son:

- En ambas categorías se observa un aumento en los meses de noviembre a enero. Esto se justifica por el periodo de rendición de las pruebas de selección universitaria y la decisión que conlleva tener buenos o malos resultados en la prueba de selección.
- La categoría de carrera profesional en los últimos años ha descendido su interés por la necesidad de personal más técnico.
- En el periodo de marzo a agosto del año 2020, ambas categorías tienen bajo nivel de búsqueda en internet debido a la pandemia.

d) **Una simplificación**

Para ayudar a los estudiantes a buscar justificaciones de la variación puede plantear las siguientes interrogantes:

- ¿Por qué en los meses de noviembre a enero aumenta la búsqueda de carreras profesionales y técnicas?
- ¿Por qué en los meses de julio y agosto la búsqueda de carreras técnicas aumenta pero no las profesionales?
- En el primer semestre del año 2020, ¿por qué la búsqueda de ambos tipos de carrera descendieron?

e) **Formas de consolidar el conocimiento o habilidades pretendidas (preguntas)**

Para consolidar el conocimiento realice las siguientes preguntas cuando el grupo diga que tiene alguna respuesta a la situación:

- ¿En qué conocimiento se fundamenta su respuesta?
- ¿Cómo llegaron a establecer por qué varió la búsqueda?
- ¿Qué elementos del gráfico consideraron para establecer su respuesta?
- ¿Existe información de su respuesta que no se obtiene del gráfico? ¿Por qué ocurre esto?

f) **Una extensión**

Observe la siguiente información sobre la matrícula del primer año según el tipo de Institución de Educación Superior (IES).

Tipo de IES	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
Universidad	154.931	153.107	149.617	153.727	162.132	166.016	167.482	160.826	149.966	147.331
Instituto Profesional	108.381	124.900	126.907	124.344	125.389	120.430	123.536	126.497	114.456	127.131
Centro de Formación Técnica	60.870	63.623	65.610	62.551	59.136	59.300	59.768	61.744	56.253	58.754

Tabla 2: Matrículas de Primer Año, periodo 2012-2021. Fuente: Elaboración propia a partir de ÍNDICES Matrículas de Educación Superior 2005-2021, CNED.

Con la información de la Tabla 2, evalúe las conclusiones que se pueden obtener del gráfico de línea.

Recomendación: Invite a los estudiantes a analizar cómo se relacionan las cifras de matrícula con las tendencias de búsqueda, promoviendo así una interpretación crítica de los datos.

g) **Plenaria**

Para la plenaria se deben seleccionar respuestas que faciliten la discusión entre los estudiantes. Las respuestas en esta situación que pueden ser elegidas deben considerar las siguientes características:

- El conocimiento en cual fundamentan su respuesta es ajeno a la información que se muestra en el gráfico. Esto para comprender y sentar las bases de un proceso de inferencia informal.
- La información externa debe ser discutible para generar asimetría en las posiciones de los estudiantes.
- Que establezcan interpolaciones/extrapolaciones, y con ello poder llegar a validar la idea en función de los elementos estructurales identificados.

2.2. Contenido para el profesor.

2.2.1. Disciplinar

La identificación de la o las **variable(s)** es fundamental para describir un fenómeno de interés y su inherente comprensión, pero desde esta perspectiva, ¿cómo se define este concepto? Intuitivamente, se asume que dicho concepto está asociado a algo que varía, o sea, a una característica a la que se le pueden asociar distintos valores o atributos. Siguiendo esta idea, consideremos que estamos interesados en analizar la siguiente hipótesis: ¿los estudiantes diagnosticados con hiperactividad tienen un menor promedio de notas en matemáticas que los estudiantes que no están diagnosticados? Características importantes de los estudiantes para llegar a conclusiones e inferencias al respecto podrían ser: el promedio de notas en la asignatura, el estudiante está diagnosticado o no con hiperactividad, la edad, el sexo, entre otras. Claramente, si tomamos una muestra representativa de la población para llevar a cabo el estudio y vemos los posibles valores o atributos de cada una de las características señaladas, verificaremos que efectivamente estas cambian de individuo a individuo. En particular, a cada una de estas características las denominaremos **variable estadística**, las cuales se subdividen en dos grandes grupos:

- **Cualitativas**, corresponden a aquellas características del sujeto de interés, que como su nombre lo indica, son una cualidad. Este tipo de variables se subdivide a su vez en dos:
  - Nominales, a este tipo de variable pertenecen todas las cualidades de los sujetos de estudio, donde los atributos no tienen un orden establecido o natural, por ejemplo:
    a) el estado civil, considerando como posibles valores de atributo: soltero, casado, viudo, otro.
    b) el color de ojos, donde los posibles valores son: negros, verdes, azules, cafés.
  - Ordinal, en este tipo de variable, los atributos tiene un orden preestablecido, el cual se debe a su naturaleza y no a la subjetividad del investigador, por ejemplo:
    a) el nivel de tolerancia al dolor, la cual puede considerar valores numéricos entre 0 y 10, donde 0 corresponde a ausencia de dolor, y 10 representa el nivel más alto (peor dolor del mundo). Note que si bien se asignaron números a las categorías de la variable, esta sigue siendo de naturaleza cualitativa.
    b) el estado nutricional, donde la variable puede tomar valores como: por debajo del peso saludable, normal, sobrepeso, obesidad, obesidad extrema.
- **Cuantitativas**, son aquellas variables en que las características del sujeto por naturaleza se expresan de manera numérica. Existen dos tipos de variables cuantitativas:
  - Discretas, corresponden a que ellas variables numéricas en que sus posibles valores son obtenidos a través del proceso de contar, por ejemplo: número de automóviles de un color específico que pasan por una intersección medida de manera diaria durante un mes.
  - Continuas, a este grupo pertenencen las variables en que sus posibles valores provienen del proceso de medir, por ejemplo: la edad (en años) de los estudiantes de un determinado establecimiento. En este ejemplo en particular, la unidad de medida establecida son los años y esta no puede tomar números decimales, sin embargo, esta no es razón suficiente para pensar que la variable en cuestión es discreta, sólo se puede señalar que por comodidad esta se discretizó, pero su naturaleza sigue siendo continua. Conclusión similar se puede establecer con otras variables continuas como el sueldo.

Note que el concepto de variable estadística está ligado a lo que se espera obtener cuando se realiza una pregunta estadística. Entenderemos que una pregunta estadística es aquella que involucra las siguientes ideas: a) implica variabilidad, es decir, la respuesta a ella no es única ni determinista, sino que varía entre los individuos de interés; b) para responderla se necesitan datos (evidencia empírica), no basta con conocimiento teórico ni lógico; c) análisis de datos, las respuestas obtenidas pueden ser resumidas, tanto numéricamente como visualmente, o a través de análisis más sofisticados como inferencias.

Preguntas como: ¿Cuántas horas tiene un día?, ¿Mañana sale el sol?, no corresponden a preguntas estadísticas, pues a pesar de que le preguntemos a 1000 personas, la respuesta es la misma, no hay variabilidad, no se necesitan datos para responderlas y no tiene sentido realizar ningún análisis de datos. Sin embargo, preguntas como: ¿Cuántas horas al día miras las redes sociales?, ¿a qué hora sale el sol en Concepción durante junio?, sí lo

son, dado que varían, en el primer caso de individuo a individuo, y en el segundo de día a día; se recogen datos empíricos, y se puede realizar un análisis de estos.

A continuación la Figura 3 muestra un resumen del tipo de variable estadística.

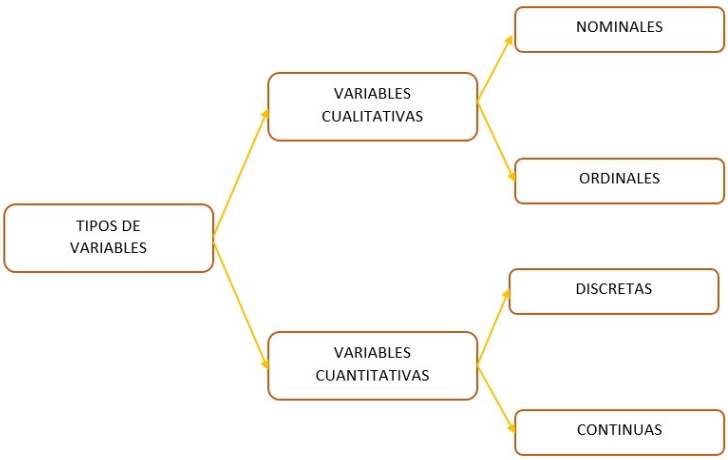


Figura 3: Tipos de variable estadística.

Una manera de resumir la información recopilada de una variable es a través de representaciones gráficas, las cuales deben tratar de retratar el comportamiento de la característica estudiada; por ende, es de vital importancia escoger el gráfico adecuado según el tipo de variable en cuestión. Por ejemplo, el **gráfico de líneas** es una forma gráfica que se construye con el objetivo de permitirle al lector estudiar y/o analizar la evolución de las observaciones relativas a una variable de interés cuantitativa, usualmente, a través del tiempo; de manera tal que este pueda, por ejemplo, proyectar la tendencia general detectada. A esta última acción la llamaremos extrapolación, asociada al nivel de lectura de gráfico más alto, tema que discutiremos en la sección siguiente.

De acuerdo con lo anterior, podemos señalar que un gráfico de líneas está formado, como su nombre lo indica, por líneas que representan datos con una estructura temporal. Para construir este tipo de gráfico se considera un plano cartesiano, donde en el eje de las abscisas (eje X) se ubica la variable relativa al tiempo (meses, trimestres, años, etc.) y en el eje de las ordenadas (eje Y) la variable de interés, la cual claramente es obtenida en instantes de tiempo discretos. Sin embargo, y dado que el tiempo es continuo y con el propósito de ayudar al lector con la lectura y análisis de la tendencia de los datos, las observaciones se unen con líneas.

El gráfico de líneas también es útil para realizar comparaciones, ya sea de dos o más variables cuantitativas versus el tiempo, por ejemplo: precio del dólar durante el último año y precio de la bencina en el mismo período (Ver Figura 4); o la misma variable medida en muestras distintas, por ejemplo: tasa de desempleo a nivel nacional según género (Ver Figura 5). Note que en la primera de estas figuras, el tiempo de observación es de un año, ambas variables presentan una tendencia al alza en dicho período. También se puede observar que el precio del dólar mostró un peak alrededor de la navidad de 2021, para luego bajar su valor; sin embargo, una nueva alza se observa en torno a la primera semana de junio del 2022. Por su parte, en el gráfico de líneas referente al precio de la bencina, se observa que este es prácticamente creciente en todo el periodo, presentando una baja alrededor de la primera semana de diciembre de 2021. Ahora bien, como ambos gráficos tienen su propia escala, su comparación no es directa.

Por otro lado, en la Figura 5 se puede realizar una comparación directa entre las tendencias de la variable en cuestión, pues no solo se usó la misma escala, sino que además se consideró graficarla en la misma ventana. El período de observación es del trimestre dic-feb 2018, hasta dic-feb de 2022, con lo cual se consideran 49 trimestres móviles. Note que el concepto trimestres móviles hace referencia a un promedio de tres períodos (en nuestro caso tres meses) consecutivos que se calcula y actualiza constantemente a medida que avanza el tiempo. Los trimestres móviles tiene la ventaja de que se ajusta constantemente a nuevos datos, lo cual permite reflejar la evolución más reciente de la variable de interés, contrario a lo que sucede con un promedio simple de tres meses que solo se calcula una vez.

Cuando el objetivo es realizar comparaciones de tendencias se recomienda que los datos sean representados

en un mismo gráfico o en gráficos distintos pero poniendo atención en las escalas de estos, para evitar distorsionar la información y con ello las conclusiones e inferencias.

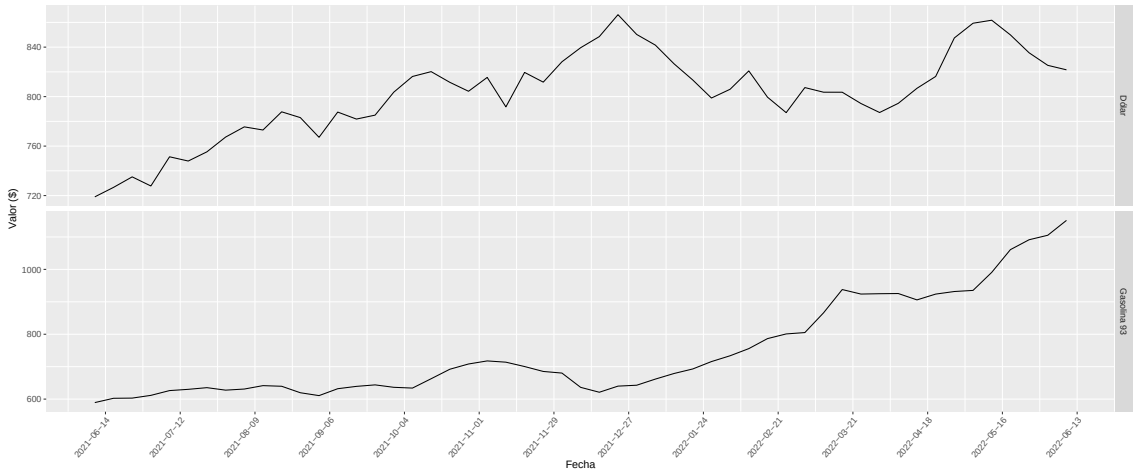


Figura 4: Precio del dólar y precio de la Gasolina de 93 octanos. Fuente: Elaboración propia en base a datos del Banco Central

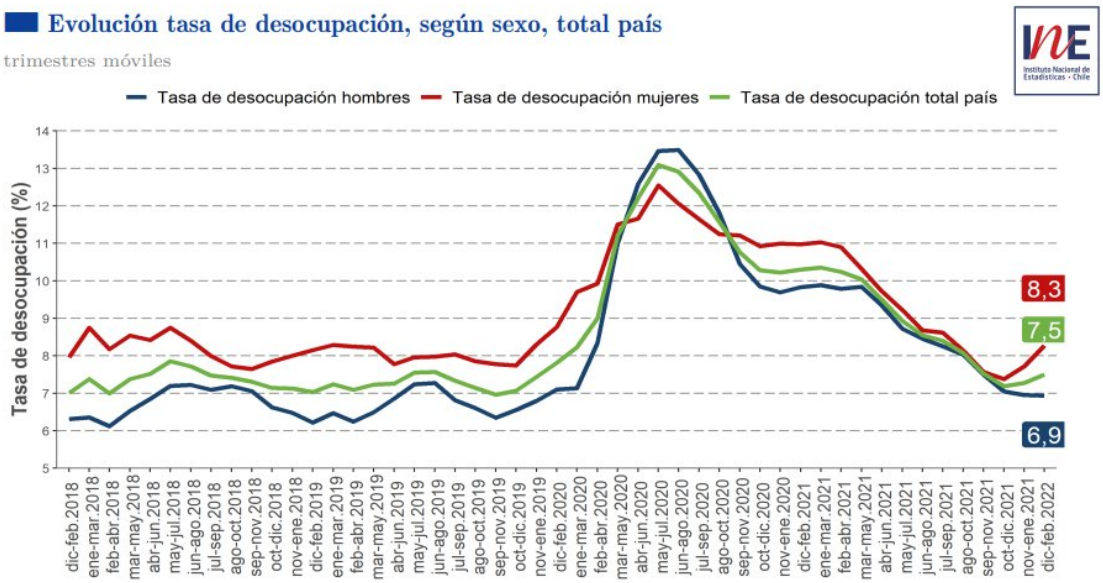


Figura 5: Tasa de desempleo a nivel nacional. Fuente: INE

2.2.2. Comprensión gráfica y razonamiento estadístico: Fundamentos cognitivos y aplicaciones didácticas

[agregar cuestiones sobre neurociencia] El desarrollo del razonamiento estadístico y la interpretación de gráficos requiere tanto de conocimientos matemáticos como de procesos cognitivos complejos. A continuación, se describen los aspectos clave que intervienen en esta comprensión, desde sus fundamentos gráficos hasta su proyección didáctica en el aula.

Comprensión de gráficos

Los gráficos transmiten información mediante una variedad de recursos visuales como puntos, líneas, sistemas de coordenadas, figuras, números, símbolos, texto y color, los cuales deben ser interpretados de manera integrada. Según Kosslyn (1985), la comprensión de un gráfico requiere considerar cuatro elementos estructurales esenciales: el fondo, la estructura, el contenido y las etiquetas. Complementariamente, Curcio (1987) destaca como elementos fundamentales las palabras presentes en el título, los ejes y las escalas, así como el contenido matemático

que incorpora —como números, áreas o longitudes de líneas—, todos ellos imprescindibles para una lectura e interpretación adecuada del gráfico.

La comprensión e interpretación de gráficos desempeña un papel central en la descripción, organización y análisis de datos estadísticos, ya que constituyen instrumentos clave en procesos de *transnumeración*. Esta última, de acuerdo con Wild y Pfannkuch (1999b), consiste en la capacidad de transformar representaciones de datos de un sistema a otro con el propósito de alcanzar una mejor comprensión. Por ejemplo, un niño que convierte un listado de datos en un gráfico de barras —y viceversa— está realizando un proceso de transnumeración que implica habilidades cognitivas complejas.

Elementos estructurales de un gráfico para la comprensión

Al solicitar a los estudiantes que analicen un gráfico, se espera que lo interpreten en relación con sus conocimientos previos y su contexto. Para ello, es fundamental que reconozcan y comprendan los elementos estructurales que lo componen. En este sentido, Friel et al. (2001) identifican los siguientes componentes esenciales de un gráfico estadístico:

- **Título y etiquetas**: Proveen información contextual sobre el contenido del gráfico e indican las variables representadas.
- **Marco del gráfico**: Incluye los ejes, las escalas y las marcas de referencia. Este marco proporciona información clave sobre las unidades de medida utilizadas y orienta la lectura cuantitativa de los datos.
- **Especificadores**: Son los elementos gráficos que representan los datos, tales como los rectángulos en los histogramas o los puntos en los diagramas de dispersión. Diversos autores han señalado que no todos los especificadores tienen el mismo nivel de accesibilidad cognitiva. Se ha propuesto un orden progresivo de dificultad en su interpretación: posición en una escala homogénea, posición en una escala no homogénea, longitud, ángulo o pendiente, área, volumen y color.

Entonces, la comprensión de la información contenida en un gráfico requiere el desarrollo de tres tipos de comportamientos, según Friel et al. (2001):

1. **Traducción:** consiste en cambiar el modo de representación de la información. Por ejemplo, implica describir verbalmente el contenido de una tabla de datos o interpretar un gráfico en términos descriptivos, destacando su estructura y los elementos visuales que lo componen.
2. **Interpretación:** implica organizar la información del gráfico según su relevancia. Este proceso permite identificar relaciones entre distintos especificadores del gráfico o entre un especificador y un eje etiquetado, con el fin de extraer conclusiones significativas a partir de los datos presentados.
3. **Extrapolación e interpolación:** consideradas como extensiones del proceso de interpretación, estas acciones requieren no solo identificar la intención comunicativa del gráfico, sino también inferir posibles consecuencias a partir de los datos. Según Ochoa (2019), extrapolar e interpolar implica señalar tendencias, anticipar comportamientos futuros o completar información ausente, combinando el análisis gráfico con conocimientos previos.

En este proceso, el estudiante no solo establece conexiones entre diferentes elementos del gráfico (interpretación), sino que también aplica conocimientos externos para proyectar o inferir información adicional (extrapolación e interpolación).

Cuestionar a los estudiantes para promover el razonamiento estadístico

Un aspecto clave para favorecer la comprensión de gráficos, desde una perspectiva cognitiva, es el tipo de cuestionamiento que realizan —o al que son expuestos— los estudiantes. Formular y responder preguntas permite movilizar procesos mentales vinculados a distintos niveles de interpretación gráfica. Diversos autores han señalado que las preguntas pueden organizarse según su nivel de complejidad: mientras las preguntas de nivel básico se centran en el contenido explícito del gráfico, las de nivel superior exigen habilidades como inferir, aplicar, sintetizar y evaluar información.

En esta línea, autores citados por Friel et al. (2001) —como Wainer (1992)— propusieron tres niveles de comprensión de gráficos en función del tipo de preguntas que se pueden plantear:

1. **Nivel elemental:** centrado en la extracción directa de datos desde el gráfico.
2. **Nivel intermedio:** enfocado en la interpolación y la identificación de relaciones entre los datos presentados.
3. **Nivel avanzado:** implica extrapolar información y analizar relaciones implícitas, promoviendo una comprensión profunda de la estructura de los datos representados.

Complementariamente, Friel et al. (2001) también propuso una clasificación basada en cuatro niveles de lectura gráfica, que detallan las formas en que los estudiantes procesan la información visual:

- **Leer los datos:** lectura literal de la información contenida en el gráfico. Por ejemplo, identificar la variable representada en el eje X.
- **Leer dentro de los datos:** lectura que requiere aplicar procedimientos matemáticos básicos (como sumas, restas o comparaciones). Un ejemplo sería calcular el rango de un conjunto de datos a partir de su valor máximo y mínimo.
- **Leer más allá de los datos:** implica realizar inferencias o predicciones que no se pueden deducir directamente del gráfico ni mediante operaciones matemáticas simples. Por ejemplo, anticipar una tendencia futura a partir del patrón observado.
- **Leer detrás de los datos:** supone una evaluación crítica del gráfico, considerando tanto el proceso de recolección y organización de datos como el contexto en el que fueron producidos. Este nivel exige conocimientos matemáticos avanzados y comprensión contextual Díaz-Levicoy et al. (2015)

De acuerdo con esta taxonomía, los estudiantes suelen desenvolverse con mayor facilidad en el primer nivel (leer los datos), mientras que los niveles superiores presentan desafíos crecientes. Las dificultades más comunes se vinculan con limitaciones en el dominio del contenido matemático, habilidades lectoras y comprensión del lenguaje gráfico. Particularmente, en el nivel *leer más allá de los datos*, los estudiantes deben hacer inferencias a partir de la representación, lo cual requiere un procesamiento cognitivo más complejo y, por tanto, representa una de las mayores dificultades en la comprensión gráfica.

2.3. Orientaciones desde la implementación

A continuación se presenta una serie de recomendaciones para la implementación del problema, las cuales han surgido desde la experiencia de la aplicación de este en el aula.

1. Activación de conocimientos previos. Esta fase se recomienda realizarla antes de mostrar el problema a resolver y comenzarla con una conversación guiada, que considere preguntas tales como:
   a) ¿Qué información pueden obtener de un gráfico de líneas?
   b) ¿Qué representa cada eje en un gráfico de líneas?
   c) ¿Qué significa que la línea resultante del gráfico suba o baje?
   d) Para contextualizar, considere ejemplos cotidianos y cercanos a los estudiantes, tales como: temperatura diaria, precio de un artículo en plataformas como Shein, Aliexpress, etc., para reforzar la noción de variación en el tiempo.
2. Exploración y lectura crítica del gráfico. Esta etapa se puede realizar para ayudar a los estudiantes en la resolución del problema planteado, pero se recomienda que esto solo sea aplicado a los grupos que están teniendo dificultades en la resolución.
   a) Pedir a los estudiantes que describan tendencias generales (aumento, disminución, estabilidad) en cada gráfico.
   b) Comparar los gráficos de universidades y de institutos profesionales, considerando preguntas como:

- ¿En qué periodos se observan cambios notables?
- ¿Se comportan de forma similar?
- Identificar los elementos estructurales: ejes, unidades, intervalos de tiempo, puntos de datos.

c) Realizar preguntas destinadas a promover la interpretación y la argumentación, por ejemplo:

- ¿Qué año tuvo el mayor número de búsquedas?
- ¿Qué podrías inferir que pasará si la tendencia continúa?
- ¿En qué año ambos gráficos muestran valores similares?

3. Conexión con la realidad y pensamiento crítico, durante el desarrollo y dependiendo del avance de los grupos, motivar la reflexión, en torno a preguntas como:

   a) ¿Qué factores sociales, económicos o educativos podrían explicar estas variaciones?

   b) ¿Qué implicancias podrían tener estos factores para la oferta educativa?

4. Actividades complementarias para reforzar contenido. Pedir a los estudiantes que creen su propio gráfico de línea a partir de una base de datos (por ejemplo, usando planillas de cálculo). Y comenten sus hallazgos, relativos a descripción de tendencias y puntos críticos.

5. Se recomienda promover una lectura crítica de los gráficos, de modo que los estudiantes puedan interpretar la información que estos presentan con base en su propio conocimiento y experiencia. Sin embargo, es importante enfatizar que ese bagaje personal no debe imponerse por sobre la evidencia mostrada en el gráfico. Asimismo, se sugiere fomentar la distinción entre correlación y causalidad, fortaleciendo así una comprensión más rigurosa y reflexiva de los datos.

3. Problema 2

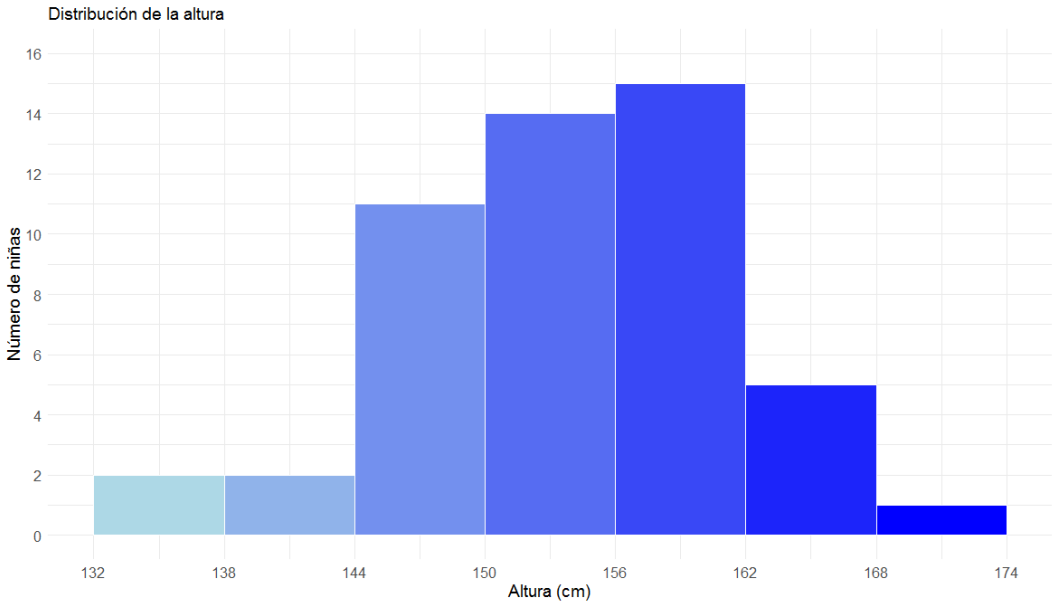
3.1. Planificación del Problema

a) **Identificación curricular y de contenido del problema**

Objetivo de Aprendizaje del Currículo Escolar	Objetivo de Aprendizaje de la Actividad	Contenidos Estadísticos y de Probabilidad
OA 1. Argumentar y comunicar decisiones a partir del análisis crítico de información presente en histogramas, polígonos de frecuencia, frecuencia acumulada, diagramas de cajón y nube de puntos, incluyendo el uso de herramientas digitales. OA a. Construir y evaluar estrategias de manera colaborativa al resolver problemas no rutinarios.	Establecer conjeturas sobre la información que se genera al relacionar diferentes representaciones de un mismo conjunto de datos.	■ Histograma ■ Cuartiles ■ Boxplot (gráfico de caja)

b) **Problema** De la montaña a la caja.

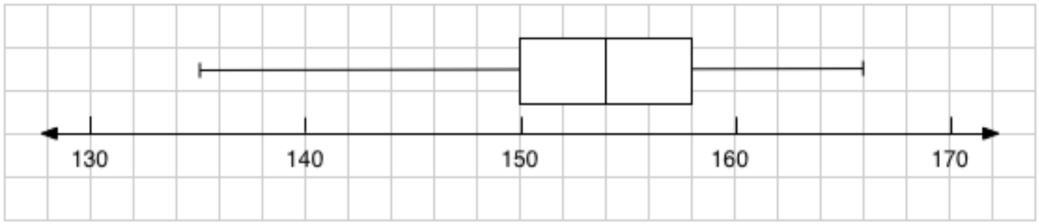
El siguiente histograma muestra la altura en centímetros de las niñas de un curso.



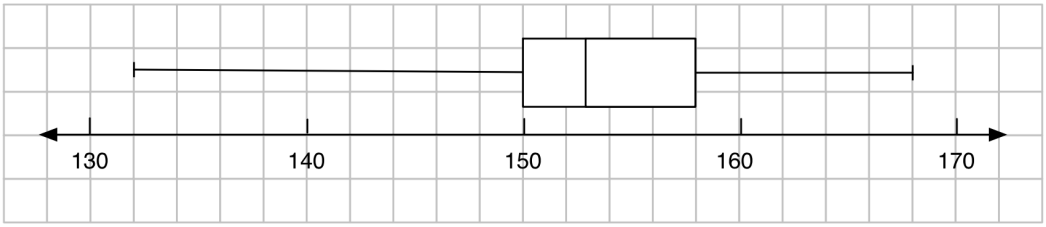
Construya el gráfico de caja asociado.

c) **Una solución del problema**

Un primer gráfico de caja solución es el siguiente:

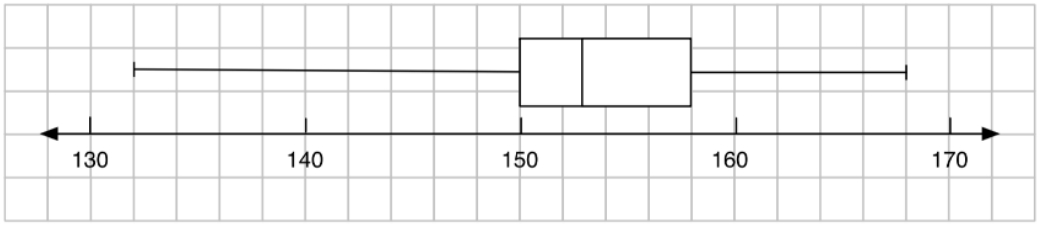


Un segundo gráfico de caja solución es el siguiente:



d) **Una simplificación**

El siguiente gráfico de caja muestra la distribución de las alturas de un grupo de niñas



Construya el histograma de la distribución de las alturas de las niñas.

e) **Formas de consolidar el conocimiento o habilidades pretendidas (preguntas)**

Para consolidar el conocimiento realice las siguientes preguntas cuando el grupo diga que tiene alguna respuesta a la situación:

- ¿Cómo saben que el valor mínimo se encuentra ahí? ¿y el máximo?
- ¿Pueden variar los valores de los cuartiles?
- ¿De qué dependen los valores del gráfico caja?
- ¿Existe otro gráfico de caja que puede ser solución?
- ¿Cuántos gráficos diferentes pueden estar asociado al gráfico de caja o al histograma o viceversa?

f) **Una extensión** Los siguientes histogramas muestran la distribución de puntajes asociados a una evaluación. Construya para cada grupo el gráfico de caja asociado.

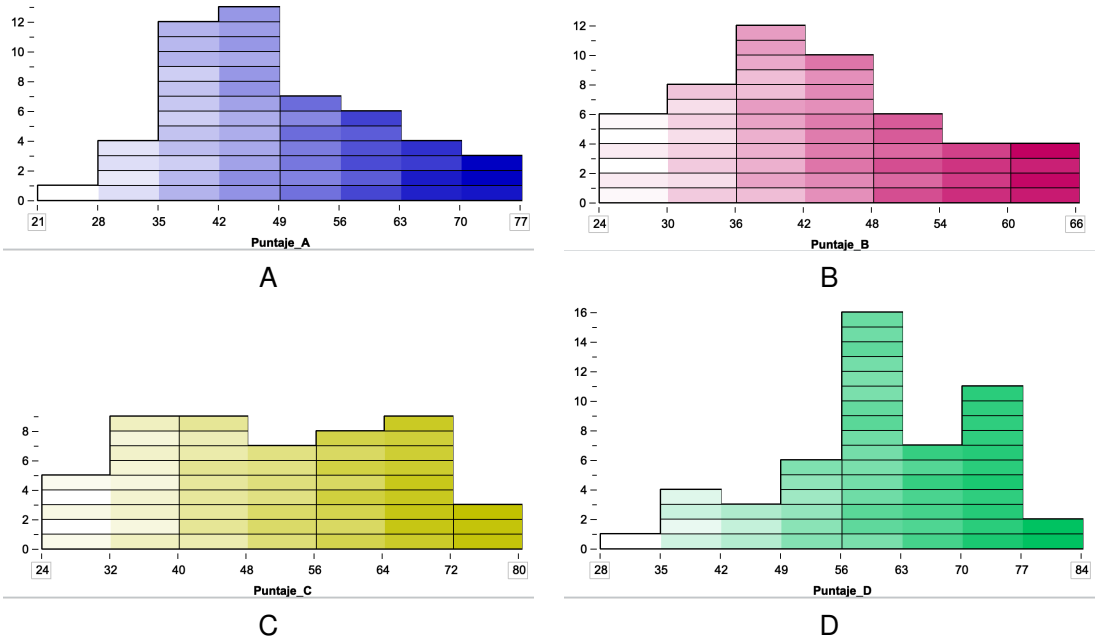


Tabla 3: Distribución de puntajes de 4 evaluaciones.

g) **Plenaria**

Para la plenaria se deben seleccionar respuestas que facilite la discusión entre los estudiantes respecto de la variabilidad de los gráficos de cajas asociados al histograma. Las respuestas que pueden ser elegidas deben considerar las siguientes características:

- Ser diferentes dos a dos los gráficos cajas creados por cada grupo de estudiantes. Esto para permitir la discusión entorno a los posibles valores que pueden tomar los cuartiles, mínimos y máximos del gráfico de caja.
- Que el conocimiento en el cual basan la variabilidad de los gráficos sea la representación agrupada de los datos en el histograma en contraposición de la necesidad de datos desagrupados para la construcción del gráfico de caja único.

Recomendación: Invite a los estudiantes a discutir sobre la presencia de datos atípicos y cómo ellos pueden influir en la forma del gráfico de caja y/o del histograma.

3.2. Contenido para el profesor.

3.2.1. Disciplinar

Organización y Representación de la Información: Una vez recopilada la información que permitirá dar respuesta a la pregunta de investigación, un paso fundamental es decidir la forma en que se organizará y presentará esta información. Para llevar a cabo esta tarea existen dos herramientas estadísticas, las cuales son complementarias: las tablas de frecuencias y las representaciones gráficas.

Antes de seguir debemos enfatizar en que el rol de una tabla de frecuencias es resumir la información recopilada, por ende es más que solo una tabla de registro, considerando que esta última consiste simplemente en una forma de disposición de la información, por ejemplo, en excel, y como tal no logra el objetivo principal de una tabla de frecuencias. Lo primero que se debe considerar en la construcción de estas tablas es el tipo de variable que se desea representar (ver Figura 3). Para variables cualitativas la tabla de frecuencias se reduce a considerar las frecuencias absolutas, relativas y relativas porcentuales. Sin embargo, cuando una variable cualitativa ordinal posee varios atributos y se requiere información adicional, se pueden obtener las versiones acumuladas.

Consideremos el siguiente ejemplo: A un grupo de estudiantes se le pregunta por su estilo musical favorito, la información recopilada se encuentra resumida en la siguiente tabla:

Estilo musical	N^o de estudiantes	$\%$ de Estudiantes
Pop	80	53,3
Reggaetón	32	21,3
Rock	13	8,7
Trap	25	16,7
Total	150	100

Tabla 4: Tabla de frecuencias del estilo musical favorito.

Note que la variable involucrada en esta situación corresponde a una variable cualitativa nominal, es decir, no se puede establecer un orden natural entre los atributos de esta. En particular, la Tabla 4, se construyó ordenando los estilos musicales alfabéticamente, considerando los 4 atributos que arrojó la variable, obtenidos de los resultados de la consulta realizada a los estudiantes. Es importante destacar que esta decisión no cambia la naturaleza nominal de la variable. La segunda columna de la tabla corresponde a las frecuencias absolutas, las cuales en sentido general representan el número de sujetos o individuos que pertenecen a cada categoría o atributo de la variable, es usual denotarlas a través de f_i, sin embargo, es importante no olvidar ¿qué representa realmente?, en nuestro caso, las frecuencias absolutas corresponden al número de estudiantes que prefieren un respectivo estilo musical. Por último, se optó por incluir la frecuencia relativa porcentual, la cual se obtiene a partir del cuociente entre la respectiva frecuencia absoluta y el total de los datos (el cual usualmente denotamos por n), por $100\,\%$. Luego si denotamos a la frecuencia relativa de la $i-$ésima categoría de la variable por r_i, tendremos que $r_i = \dfrac{f_i}{n}$, y por lo tanto, la respectiva frecuencia relativa porcentual corresponde a $r_i * 100\,\%$.

Luego, podemos deducir características importantes, tanto de las frecuencias absolutas, como de sus versiones relativas, como por ejemplo:

- La suma de las frecuencias absolutas es igual al total de datos, es decir,

$$\sum_{i=1}^{k} f_i = f_1 + f_2 + ... + f_k = n,$$

  donde k corresponde al número de categorías de la variable, en nuestro ejemplo $k = 4$.

- Las frecuencias relativas, r_i, están siempre acotadas entre 0 y 1, y su suma es igual a 1,

$$\begin{aligned}\sum_{i=1}^{k} r_i &= r_1 + r_2 + ... + r_k \\ &= \frac{f_1}{n} + \frac{f_2}{n} + ... + \frac{f_k}{n} \\ &= \frac{f_1 + f_2 + ... + f_k}{n} \\ &= \frac{n}{n} \\ &= 1\end{aligned}$$

  Es importante señalar que en ocasiones y debido al redondeo de cifras la suma no es exactamente 1, pero dado que la frecuencia relativa corresponde a una fracción, no se debe perder ese sentido de interpretación, es decir, que fracción del total de sujetos cumple con el $i-$ésimo atributo de la variable.

- Las frecuencias relativas porcentuales $r_i\,\%$, como consecuencia del punto anterior, suman 100 $\%$.

A partir de la Tabla 4 se pueden responder preguntas como: ¿a cuántos de los estudiantes entrevistados les gusta el Trap?, ¿qué porcentaje de los entrevistados prefieren el Pop? Sin embargo, preguntas como, ¿a cuántos de los entrevistados les gusta como máximo el Rock?, no tienen sentido por la naturaleza nominal de la variable. Analicemos el siguiente ejemplo:

Con el propósito de seleccionar a los usuarios de prestaciones sociales, en Chile a partir del año 2016, se implementó el Registro Social de Hogares (RSH), lo que permite una clasificación socioecómica de los usuarios que completan dicho cuestionario. De acuerdo con la Tabla 5 dicha cifra alcanzaba en julio de 2017 al 73.1 $\%$ de la población chilena. El tramo de clasificación socioeconómica (CSE) está dividido en 7 categorías, donde el Tramo del 40, es el más bajo y a él pertenecen todos los hogares con menores ingresos, mientras que en el Tramo 100 se ubican los hogares de mayores ingresos. Claramente entonces la variable Tramo CSE, corresponde a una variable cualitativa ordinal.

Hogares y Personas en el RSH según tramo de Calificación Socioeconómica. Julio 2017 (Número y Porcentaje)

	Hogares		**Personas**		**Distribución Sobre Total País (%)**
Tramo CSE	**N**	**%**	**N**	**%**	
Tramo del 40	2.654.031	55,1	7.024.262	54,8	40,0
Tramo del 50	462.475	9,6	1.306.436	10,2	7,4
Tramo del 60	374.917	7,8	1.017.426	7,9	5,8
Tramo del 70	358.587	7,4	963.631	7,5	5,5
Tramo del 80	343.064	7,1	892.471	7,0	5,2
Tramo del 90	451.636	9,4	1.234.199	9,6	7,0
Tramo del 100	168.817	3,5	388.367	3,0	2,2
Total	**4.813.527**	**100,0**	**12.826.792**	**100,0**	**73,1**

Tabla 5: Tabla de frecuencia de los hogares y personas con registro social de hogares (RSH) a nivel nacional. Fuente:Ministerio de Desarrollo Social y Familia (2017, p. 169)

Si consideramos, por ejemplo solo a los hogares, y suponemos que una determinada ayuda social se entregará a quienes pertenezcan a lo más el tercer tramo, podemos construir una tabla de frecuencias que

además permita responder a preguntas como: ¿cuántos hogares pertenecen a los más a dicho tramo?, o ¿qué porcentaje de hogares está clasificado hasta el Tramo del 60? Preguntas como estas son posibles de responder con la frecuencia absoluta acumulada, y sus respectivas versiones relativas, en cuyo caso la tabla correspondiente estará dada por:

Tramo CSE	N^o de hogares	$\%$ de Hogares	N^o de hogares	$\%$ de Hogares
Tramo del 40	2.654.031	55,1	2.654.031	55,1
Tramo del 50	462.475	9,6	3.116.506	64,7
Tramo del 60	374.917	7,8	3.491.423	72,5
Tramo del 70	358.587	7,4	3.850.010	80,0
Tramo del 80	343.064	7,1	4.193.074	87,1
Tramo del 90	451.636	9,4	4.644.710	96,5
Tramo del 100	168.817	3,5	4.813.527	100,0

Tabla 6: Tabla de frecuencias de los hogares chilenos por RSH.

Note que en la Tabla 6, las columnas 2 y 3 corresponden a las frecuencias absolutas y a las frecuencias relativas porcentuales para cada tramo CSE, mientras que las columnas 4 y 5, son las versiones acumuladas de dichas frecuencias, las que denotaremos como F_i y $R_i\,\%$, respectivamente, sin embargo, ellas representan el número y porcentaje de hogares que se acumulan hasta cumplir con un atributo en particular, respectivamente. Por lo tanto, el cálculo de estas frecuencias nos permiten contestar a las preguntas planteadas previamente. Estas frecuencias son posibles de calcular dada la naturaleza ordinal de la variable, sumado a que esta presenta 7 posibles valores distintos.

En el caso de las variables cuantitativas, la tabla de frecuencias se puede elaborar utilizando tanto las frecuencias absolutas como las acumuladas. Sin embargo, para este tipo de variables existe un formato adicional: la tabla que organiza los datos en intervalos. Por ello, antes de construir la tabla de frecuencias, es necesario responder la siguiente pregunta: ¿cuándo conviene construir intervalos para un conjunto de datos de una variable cuantitativa? La respuesta depende de la naturaleza de los datos, es decir, de cuántos valores distintos puede tomar la variable y de qué tan dispersos se encuentran. Cuando la simple consideración de clases individuales no permite resumir adecuadamente la información, se recurre a intervalos. Es importante no basar esta decisión únicamente en criterios como el tamaño de la muestra o si la variable es continua. Analicemos los siguientes ejemplos:

a) La Tabla 7 resume los datos correspondientes a la edad (en años) de los alumnos que cursaron Enseñanza Media Científico-Humanista diurna durante el año 2021, en alguno de los establecimientos pertenecientes al Servicio Local de Educación Pública (SLEP) Andalién Sur.

Edad	N^o de alumnos	$\%$ de alumnos	N^o de alumnos	$\%$ de alumnos
14	340	9,2	340	9,2
15	442	11,9	782	21,1
16	915	24,7	1697	45,8
17	1199	32,3	2896	78,1
18	538	14,5	3434	92,6
19	220	5,9	3654	98,5
20	46	1,2	3700	99,8
21	7	0,2	3707	99,9
22	2	0,1	3709	100,0

Tabla 7: Distribución de las edades de los estudiantes. Fuente: Elaboración propia a partir de **?**

Es claro que en este ejemplo la variable, edad (en años) de los estudiantes que cursaron Enseñanza Media diurna es una variable cuatitativa continua, sin embargo, como ella sólo asume 9 valores posibles (entre 14 y 22 años), no es necesario la construcción de intervalos. Incluso considerando que la cantidad de datos $n = 3709$, es un valor grande, sólo se construyeron clases individuales.

b) Por otro lado, si consideramos la edad en años de todos los estudiantes de la comuna de Concepción que cursaron tanto Enseñanza Básica como Enseñanza Media diurna, durante el año 2021, los valores fluctúan entre 6 y 22 años, por lo tanto considerar clases individuales no parece razonable para resumir la información. La Tabla 8, muestra la tabla de frecuencias construida a partir de la información disponible.

Clase	Edad	N^o de alumnos	$\%$ de alumnos	N^o de alumnos	$\%$ de alumnos
1	[5- 8]	5178	19,01	5178	19,01
2	]8- 11]	6149	22,57	11327	41,58
3	]11- 14]	6527	23,96	17854	65,54
4	]14- 17]	8157	29,94	26011	95,48
5	]17- 20]	1222	4,49	27233	99,97
6	]20- 23]	8	0,03	27241	100,0

Tabla 8: Distribución de las edades de los estudiantes. Fuente: Elaboración propia a partir de Ministerio de Educación de Chile (2022)

A partir de estos ejemplos, es claro que la construcción de intervalos es una decisión que debe tomar el investigador de acuerdo a cómo son sus datos en relación a la cantidad de valores posibles que tome su variable, y no está ligada a la naturaleza continua o discreta de la variable ni a la cantidad de datos que se disponga. Luego, si la decisión involucra la creación de intervalos, la pregunta natural que surge es ¿cuántos intervalos debo contruir? La respuesta no es simple, sin embargo, existen varias reglas que han ido surgiendo con los años para ayudar en esta labor, las más utilizadas son:

- **Método empírico:** esta regla depende del criterio del investigador, por lo tanto es arbitrario, debido a que él decide la cantidad de intervalos que desea construir considerando:

$$5 \leq \kappa \leq 20,$$

  donde κ representa el número de intervalos.
- **Raíz cuadrada:** esta regla indica que para calcular el número de intervalos sólo se debe obtener la raíz cuadrada del total de datos, es decir,

$$\kappa = \sqrt{n},$$

  Hirai (1989) señala que este criterio es simple.
- **Regla de Sturges:** Helbert Sturges, propone que para determinar el número de intervalos se debe considerar la siguiente expresión (Sturges, 1926):

$$\kappa = 1 + 3,322\text{log}(n),$$

  expresión que surge de considerar la aproximación de la distribución binomial a la distribución normal y el supuesto que la dispersión relativa aumenta logarítmicamente.

La pregunta ahora es, ¿cuándo es conveniente uno u otro criterio?, analicemos esto de manera gráfica. En la Figura 6 se observa que para un tamaño muestral aproximadamente $n = 40$ los criterios en cuestión coinciden, mientras que para un valor menor a 40 parece ser que el criterio de usar la raíz proporciona una cantidad menor de intervalos, lo contrario ocurre cuando $n > 40$. Sin embargo, se debe tener en consideración que dado que la cantidad de intervalos debe ser un número entero, en ambos criterios de deben redondear los resultados, a pesar de ello parece ser que dado su simplicidad es aconsejable usar el criterio de la raíz para tamaños muestrales iguales o menores a 40 (incluso considerando que la cantidad de intervalos arrojada por Sturges coincide al efectuar el redondeo respectivo).

En la Figura 7, se observa como se comportan estos criterios al considerar el redondeo al entero más cercano, note que esto también es subjetivo, porque también la decisión podría ser truncar dicha cantidad. En la figura en cuestión se observa que efectivamente por simplicidad entorno a $n = 40$ es preferible usar el criterio de la raíz cuadrada.

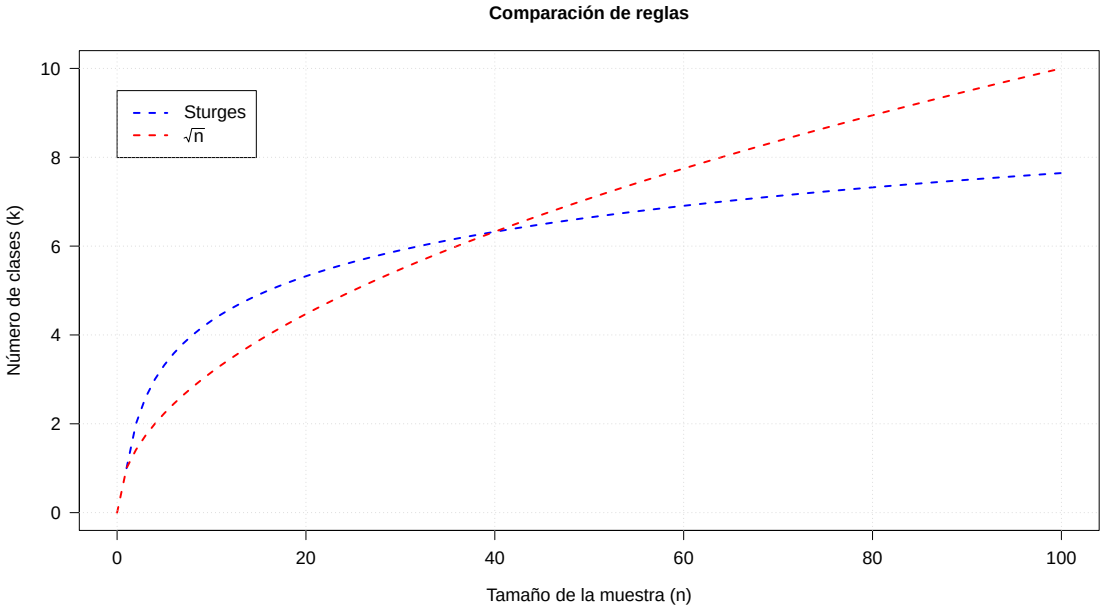


Figura 6: $\sqrt{n}$ versus Regla de Sturges (sin redondeo).

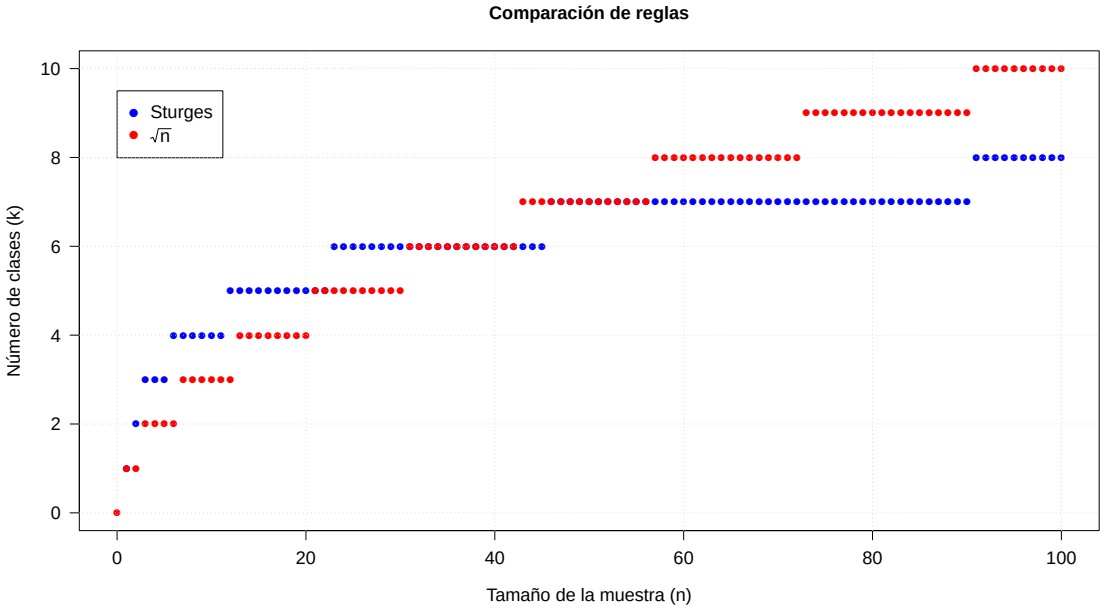


Figura 7: $\sqrt{n}$ versus Regla de Sturges (con redondeo).

Aplicando lo anterior a los datos de la Tabla 8, y considerando que $n = 27241$, utilizando la regla de Sturges se obtiene: $\kappa = 1 + 3,322 \log(27241) = 15,7338 \approx 16$ intervalos, es decir, para este conjunto de datos de acuerdo con esta regla deberían haberse construido 16 intervalos. Lo cual es ilógico al observar los datos, debido a que los valores de la variable fluctúan entre 6 y 22 años, luego es prácticamente equivalente a considerar clases individuales, razón por la cual se optó por el método empírico, es decir, se decidió a conveniencia la cantidad de clases.

Otro elemento fundamental en la construcción de intervalos es la amplitud de estos, a la que denotaremos por a, para ello básicamente existen dos opciones: el investigador lo decide de manera arbitraria y

a conveniencia, o se utiliza la siguiente expresión:

$$a = \frac{R}{\kappa},$$

donde R representa el rango de los datos, el cual se obtiene a partir de la diferencia entre el dato máximo y el dato mínimo. Se recomienda que a sea redondeado de acuerdo a la naturaleza de la variable, es decir, si los valores que asume la variable son enteros, la amplitud debe aproximarse a un entero, si los datos consideran un decimal, a también debe ser aproximado a la décima, y así sucesivamente. Por simplicidad, todos los intervalos son construidos considerando la misma amplitud.

Por último, se debe decidir con respecto a los límites de los intervalos, a los cuales denominamos como límite inferior y superior, respectivamente, los cuales pueden ser cerrados y/o abiertos, decisión que toma el investigador. Note que en la Tabla 8 se consideró como primer límite inferior a 5 años, un valor no observado, mientras que su respectivo límite superior es 8 años que si corresponde a un valor observado de la variable. En particular esta tabla fue construida a conveniencia, donde el primer intervalo es cerrado tanto inferior como superiormente, y los restantes son abiertos inferiormente y cerrados superiormente. Y la amplitud considerada es 3 años.

Sin embargo, independiente del método utilizado y las decisiones que se tomen para construir los intervalos, deben cumplirse dos condiciones fundamentales: el primer intervalo debe contener al valor mínimo y el último intervalo al valor máximo, es decir el primer y último intervalo no pueden tener frecuencia nula. Y la segunda condición es que los intervalos deben ser excluyente, la razón es muy obvia, cada dato sólo puede estar contemplado en un único intervalo de clase.

A las tablas que consideran construcción de intervalos también se les llama tabla de frecuencias con datos agrupados.

Otra manera de organizar y resumir la información es a través de la construcción de gráficos, en particular discutiremos tres tipos de gráficos: gráfico de barras, histograma y boxplot. Las representaciones gráficas son consideradas una manera de complementar la información proporcionada por las tablas de frecuencias.

- **Gráfico de barras simples:** Este tipo de gráfico se utiliza tanto para variables cualitativas como para variables cuantitativas, salvo en el caso de que estas últimas se encuentren agrupadas en intervalos. Para su construcción, en el eje X se representan los valores o categorías que puede asumir la variable, mientras que en el eje Y se ubican las frecuencias correspondientes, las cuales determinan la altura de cada rectángulo. Para que el gráfico cumpla adecuadamente su función, es importante incluir elementos estructurales como los nombres de los ejes y un título que facilite la interpretación de la información por parte del lector. Veamos algunos ejemplos:
  a) Si consideramos el ejemplo correspondiente a la Tabla 4, donde la variable era de tipo cualitativo nominal, el gráfico de barras respectivo está dado por la Figura 8, en ella se observa que cada rectángulo construido tiene como altura el número de estudiantes que prefieren cada estilo musical, es decir, en este caso se graficó la frecuencia absoluta. Debido a la naturaleza cualitativa de la variable, este gráfico también puede elaborarse de manera vertical. Note que en particular como esta variable tiene sólo 4 categorías (o estilos musicales), un gráfico circular o de torta, puede ser más útil para representar la información, en cuyo caso se utilizaría la frecuencia relativa porcentual.
  b) La Figura 9, representa el gráfico de barras respectivo a la Tabla 7, donde la variable de interés corresponde a la edad (en años) de los estudiantes que cursaron la Enseñanza Media durante 2021, en establecimientos pertenecientes al SLEP Andalién Sur, la cual es de tipo cuantitativa continua.
  Los dos ejemplos anteriores muestran que el gráfico de barras simples se puede elaborar, tanto para variables cualitativas, como para variables cuantitativas, donde la tabla de frecuencias de estas últimas contempla sólo clases individuales.
- **Histograma:** este gráfico se construye igual que el gráfico de barras, es decir, se utilizan rectángulos de altura igual a las frecuencias que se quieren representar. Sin embargo, en esta representación las barras deben ir juntas, debido a que se construye a partir de tablas de frecuencias con datos agrupados,

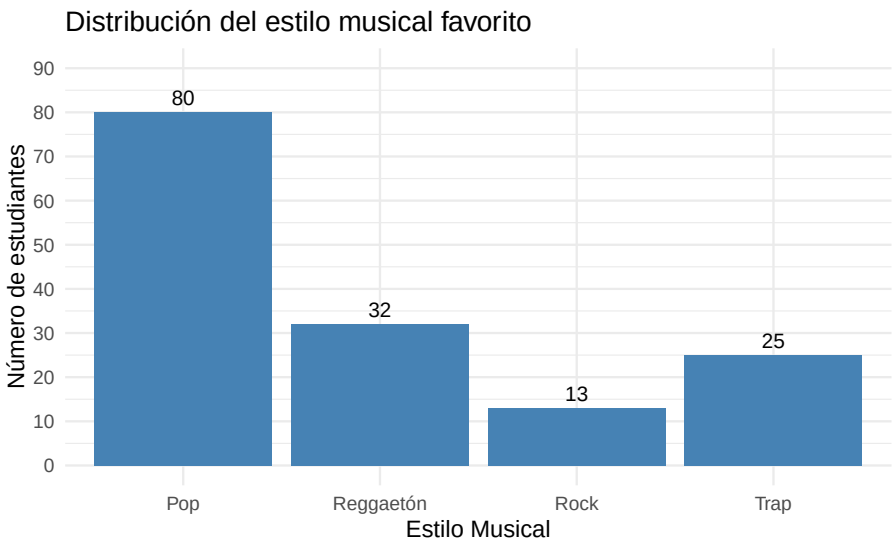


Figura 8: Gráfico de barras estilo musical favorito de un grupo de estudiantes.

Figura 9: Gráfico de barras de la distribución de las edades de los estudiantes que cursaron Enseñanza Media durante el año 2021 en establecimientos del SLEP Andalién Sur.

en cuyo caso se debe llenar toda la escala horizontal, dado a que esta representa la recta real. De acuerdo a lo anterior, sólo es posible elaborar el histograma cuando se esté tratando con una variable cuantitativa cuya tabla de frecuencias contemple la construcción de intervalos, es decir, datos agrupados. Note que el eje X, tanto en el gráfico de barras simples como en el histograma, puede empezar en un valor a conveniencia, es decir, no necesariamente, la escala del gráfico comienza en el origen del plano cartesiano.

La Figura 10, corresponde a la representación gráfica mediante histograma de la Tabla 8, note que en el eje X se consideran los intervalos construidos, sin embargo, es común encontrar histogramas donde las etiquetas respectivas de las barras sean las marcas de clases, las cuales se obtienen como el promedio de los respectivos límites.

- **Boxplot:** el boxplot, gráfico de caja o gráfico de bigotes, es una representación gráfica cuyo principal objetivo es entregar información sobre la posible asimetría de los datos, ideas intuitivas sobre dispersión de estos, y analizar la presencia de observaciones atípicas (outliers) o extremas. Este gráfico sólo se puede realizar para variables cuantitativas. Para lograr lo anterior se deben considerar varias pasos preestablecidos:

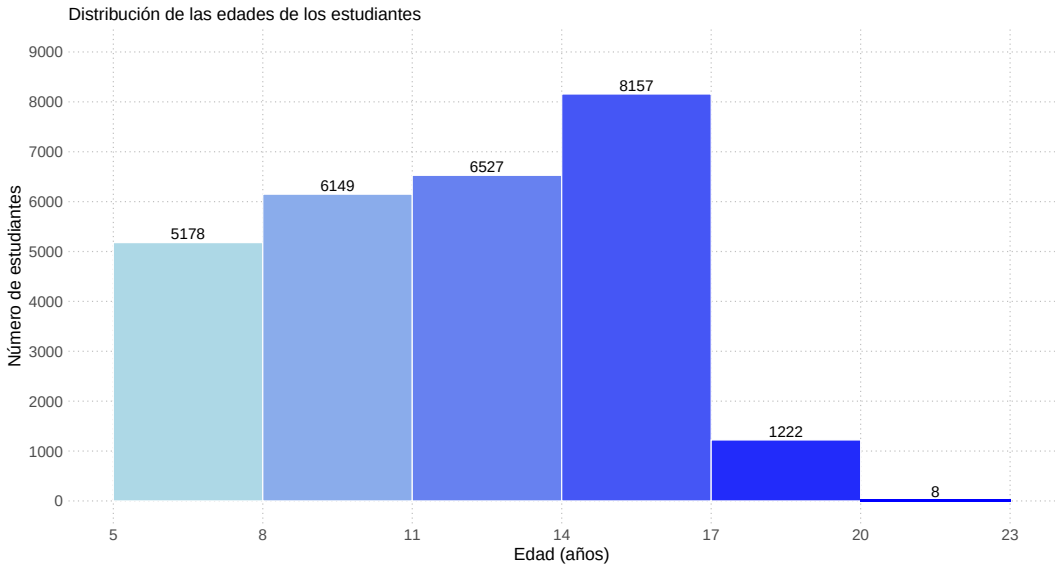


Figura 10: Histograma de la distribución de las edades de los estudiantes que cursaron Enseñanza Básica y Media en establecimientos del SLEP Andalién Sur, durante el año 2021.

- Identificar el valor mínimo y valor máximo de los datos.
- Calcular los cuartiles, los cuales corresponden a un tipo de cuantil en particular, que dividen al conjunto de datos en 4 partes de igual tamaño (misma cantidad de datos), y ayudan a describir la distribución de los datos. Por lo cual, diremos que existen tres cuartiles a los que denotaremos como Q_1, Q_2 y Q_3, respectivamente, el cuartil 2 en particular corresponde a la mediana. Para poder calcularlos, los datos deben estar previamente ordenados en sentido creciente (de menor a mayor).

  Para calcular los cuartiles lo usual es considerar la posición de cada uno de ellos en el conjunto de datos, las que se obtienen de acuerdo con:

  $$Pos(Q_1) = (n+1) * 0,25, \quad Pos(Q_2) = (n+1) * 0,50 \quad \text{y } Pos(Q_3) = (n+1) * 0,75,$$

  donde recordemos n corresponde al total de datos que disponemos. Note que de estas expresiones se pueden obtener resultados enteros o decimales, por lo cual se deben tomar decisiones al respecto y establecer convenciones. La convención más común es extender la usada para calcular la mediana, la que indica que si la posición respectiva es un número decimal, se debe tomar el promedio entre las dos observaciones entre dicho número. Por ejemplo, si se dispone de $n = 26$, la posición relativa al primer cuartil sería: $Pos(Q_1) = 27 * 0,25 = 6,75$, es decir, el cuartil 1, sería el promedio de los datos ubicados en la posición 6 y 7, respectivamente. Con ello note que los cuartiles no necesariamente son un valor observado de los datos. Si bien existen varias convenciones para calcularlos, lo importante es su intepretación, la cual corresponde a afirmar por ejemplo: Sea $Q_1 = a$ el cuartil a intepretar, entonces, diremos que "al menos el $25\,\%$ de los datos en cuestión (de las observaciones) son menores o iguales a a, mientras que al menos el $75\,\%$ de ellas es mayor o igual a a". Del mismo modo se interpretan los restantes dos cuartiles.
- Dibujar la caja, la cual corresponde a un rectángulo de largo la medida del rango intercuartílico, $IR = Q_3 - Q_1$. Esta cantidad corresponde a una medida de dispersión de los datos, que concentra el $50\,\%$ de las observaciones, es menos sensible que el rango, R, a la presencia de datos atípicos.
- Ubicar la mediana (o Q_2) en la caja. Ubicar la media o promedio. Si bien estos conceptos son tratados como sinónimos, debemos destacar que la media corresponde a un parámetro (se mide en la población), mientras que el promedio es un estimador de la media (se mide en una muestra). Para calcularla existen dos fórmulas:

- Datos individuales: $\bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$, donde x_i representa cada una de las observaciones.
- Datos agrupados en clases individuales o en intervalos: $\bar{x} = \dfrac{\sum_{i=1}^{k} x_i f_i}{n}$, donde k representa el número de clases o intervalos construidos. Cuando los datos han sido agrupados en clases individuales x_i representa los valores de la variable en análisis, y si los datos están dispuestos en intervalos de clase, x_i corresponde a la marca de clase. En ambos casos, f_i es la frecuencia absoluta de la clase en cuestión. Note que en particular para datos agrupados en intervalos, la fórmula señalada entrega una aproximación del promedio, debido a que en ella actúa la marca de clase, la recomendación entonces es que sí se dispone de los datos que originaron la tabla de frecuencia en cuestión, sean estos los utilizados para obtener el promedio.

El promedio sólo se puede obtener para variables cuantitativas, independiente de si esta es discreta o continua. Es importante señalar que a pesar que los atributos de una variable cualitativa sean expresados de manera númerica no es posible calcular el promedio, debido a que la variable sigue siendo de naturaleza cualitativa. Por ejemplo, se le pide a un grupo de personas que indique su estado civil, escribiendo 1 si es soltero, 2 si es casado, 3 si es separado, etc., note que si bien la persona indicará con un número su estado civil, la variable en cuestión seguirá siendo de tipo cualitativa nominal, y por ende el cálculo del promedio carece de sentido.

Antes de seguir es importante indicar que el promedio es un valor único que numéricamente está entre el valor mínimo y máximo de los datos, pero no necesariamente es un dato observado. Incluso del cálculo respectivo se puede obtener que el promedio contiene cifras decimales, a pesar por ejemplo que la variable sea de tipo cuantitativa discreta, no obstante lo importante es su interpretación. Es decir, se deben tener en consideración que: a) intuitivamente el promedio hace alusión al concepto de reparto equitativo, precisamente esta idea permite establecer su forma de cálculo, b) el promedio es el único valor que cumple con la condición de que la suma de las diferencias del promedio a cada dato menor a él, es igual a la suma respectiva de las diferencias del promedio a cada datos mayor a él, c) y como consecuencia de lo anterior, es el único valor que cumple con que la suma de los valores absolutos de las desviaciones es cero, es decir:

$$\sum_{i=1}^{n} |x_i - \bar{x}| = 0.$$

Note que de b) se puede deducir la idea intuitiva de equilibrio que posee la media, pues $|x_i - \overline{x}|$, cuando $x_i < \overline{x}$, es igual a dicha expresión evaluado en los valores de la variable mayores al promedio ($x_i > \overline{x}$), analicemos esto gráficamente.

Como podemos ver en la Figura 11 la suma de las desviaciones de los datos menores a la media (líneas rojas), y la suma de las desviaciones correspondientes de las observaciones mayores (las líneas azules), en valor absoluto son iguales. Claramente esta figura fue elaborada de manera tal que cada dato tenía frecuencia 1 sólo con el propósito de representar la idea de manera fácil, sin embargo, esto no tiene porque ser así. Analicemos el siguiente ejemplo:

Los datos de la Tabla 9 representan el puntaje en escala 0 a 100, en una determinada prueba obtenidas por un grupo de 126 estudiantes.

Puntaje	10	20	30	40	50	60	70	80	90	100
Número de estudiantes	1	3	7	8	12	50	31	8	5	1

Tabla 9: Distribución del puntaje

Luego, se calcula el puntaje promedio obtenido por los estudiantes utilizando la expresión para

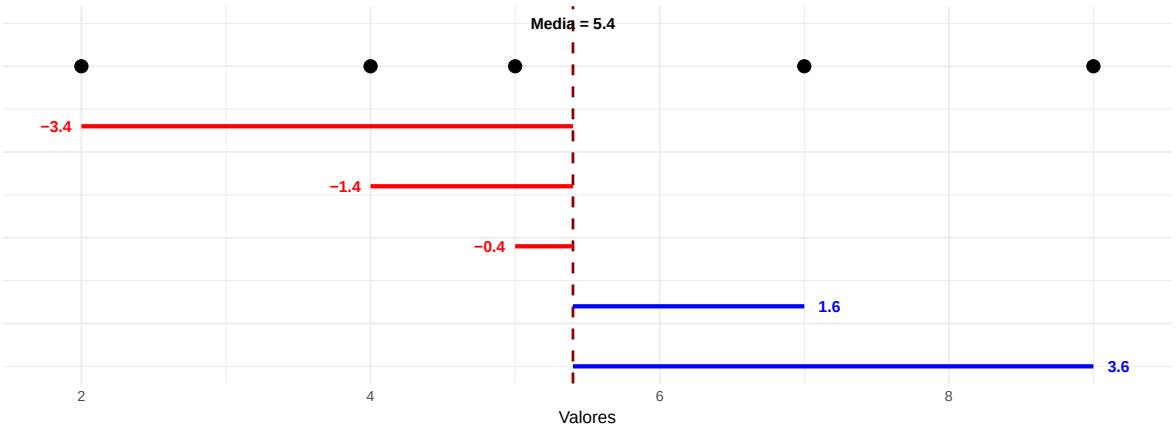


Figura 11: Idea intuitiva de equilibrio de la media.

datos agrupados en clases individuales, dadas por: $\overline{x} = \sum_{i=1}^{k} x_i f_i / n$, realizando las sustituciones respectivas resulta $\overline{x} = 60$ puntos. Entonces, note que las desviaciones de cada valor de la variable al promedio estará dada por: $(x_i - \overline{x}) \times f_i$, entonces

Puntaje	10	20	30	40	50	60	70	80	90	100
Número de estudiantes	1	3	7	8	12	50	31	8	5	1
$(x_i - \overline{x}) \times f_i$	-50	-120	-210	-160	-120	0	310	160	150	40

Tabla 10: Distribución del puntaje y sus desviaciones respecto al puntaje promedio.

Es claro de la Tabla 10 que la suma de las desviaciones de los valores de la variable menores al promedio en valor absoluto ($|-50-120-210-160-120| = 660$), es igual a la suma correspondiente de las desviaciones de los valores mayores ($|310+160+150+40| = 660$). Note que por simplicidad se propuso un ejemplo donde el promedio fuera un número entero, sin embargo, esta propiedad se cumple cualquiera sea el valor del promedio.

En cuanto a la mediana (Me), recordemos que esta es equivalente al cuartil 2, Q_2, y por ende representa el valor que divide al conjunto de datos en dos partes iguales, es un valor único, y no necesariamente es un valor observado. Este valor es menos sensible que la media o promedio ante la presencia de datos atípicos y/o asimetría de los datos, por eso es importante establecer una relación entre ellas:

- Si los datos son aproximadamente simétricos, entonces la media y la mediana son valores muy cercanos.
- Si los datos presentan asimetría a la derecha, es decir, la masa de la distribución se encuentra más concentrada a la izquieda y se observa una cola hacia la derecha, entonces, se dice que los datos tienen asimetría positiva, y por ende $Me < \bar{x}$.
- Si los datos presentan asimetría a la izquierda, se espera un comportamiento de la masa de distribución inverso a lo expresado en el punto anterior, entonces, diremos que los datos tienen asimetría negativa, y por ende $Me > \bar{x}$.

La Figura 12 ilustra estas ideas. Note que las líneas negras corresponden a una curva a mano alzada, la cual es usual que reciba el nombre de polígono de frecuencia, tema que trataremos con más detención en el capítulo siguiente.

- Representar los bigotes, los cuales corresponden a líneas que se dibujan a partir de la caja, tanto a la izquierda como hacia la derecha, sin embargo es necesario hacer algunas precisiones. En la literatura es frecuente encontrar definiciones como los bigotes o antenas son “líneas que conectan

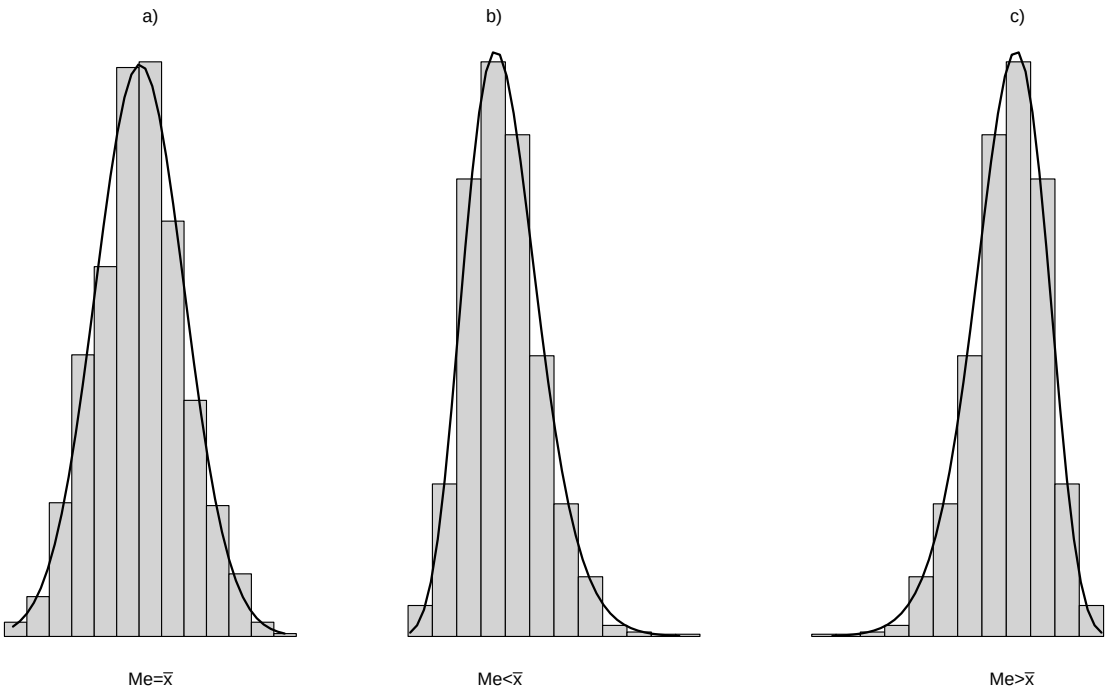


Figura 12: a) simetría, b) asimetría positiva y c) asimetría negativa.

los valores mínimo y máximo al rectángulo central"(Catalán et al., 2017, p. 318). Esta definición, aunque intuitiva, puede inducir a errores de interpretación, ya que la distribución de los datos puede presentar colas largas debido a uno o más valores extremos. Si se dibujan los bigotes directamente hasta los valores mínimo y máximo, se corre el riesgo de ocultar información relevante sobre estos valores atípicos. Por tal razón, antes de dibujar los bigotes proponemos el uso de las barreras internas (BI) y externas (BE), las cuales permiten analizar la presencia de valores atípicos o extremos. La forma de cálculo para obtener estas barreras es:

$$BI_1 = Q_1 - 1,5RI, \;\; BI_2 = Q_3 + 1,5RI,$$

$$BE_1 = Q_1 - 3RI, \;\; BE_2 = Q_3 + 3RI.$$

Esta forma de calcular las barreras, tanto internas como externas proviene de los trabajos de **?**, quien construyó una regla para identificar valores atípicos en datos con distribuciones simétricas y unimodales (con forma de campana).
Para entender de manera intuitiva las fórmulas de las barreras consideremos la Figura 13, donde $X \sim N(\mu, \sigma^2)$, a partir de ella podemos establecer las siguientes relaciones con los cuartiles 1 y 3, respectivamente.

$$\begin{array}{rrcll} & P(X \leq Q_1) & = & 0,25 & \text{Definición de } Q_1 \\ & P\left(Z \leq \dfrac{Q_1 - \mu}{\sigma}\right) & = & 0,25 & \text{Estandarizando la variable } X \\ \Rightarrow & \dfrac{Q_1 - \mu}{\sigma} & = & -0,675 & \text{Valor del percentil correspondiente con } Z \sim N(0,1) \\ \Rightarrow & Q_1 & = & \mu - 0,675\sigma & \text{Despejando adecuadamente} \end{array} \tag{1}$$

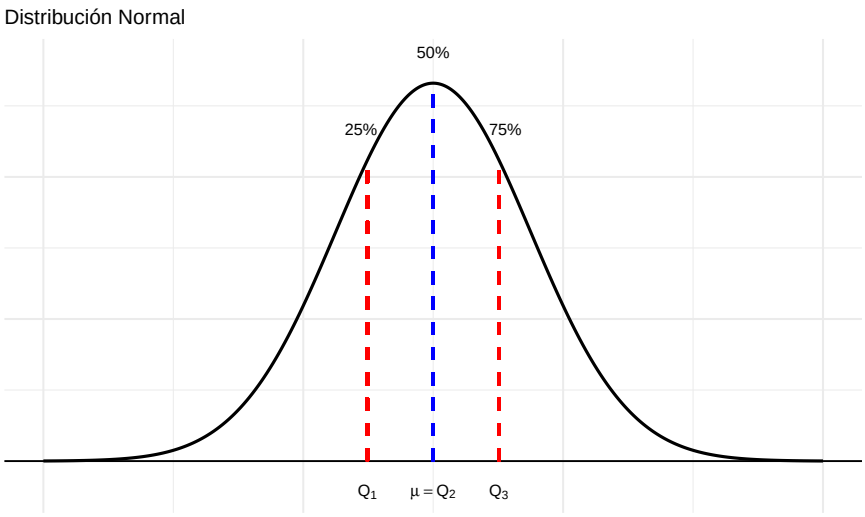


Figura 13: Posición de los cuartiles de una distribución normal.

Análogamente para Q_3 se obtiene:

$$\begin{aligned} P(X \leq Q_3) &= 0,75 && \text{Definición de } Q_3 \\ P\left(Z \leq \frac{Q_3 - \mu}{\sigma}\right) &= 0,75 && \text{Estandarizando la variable } X \\ \Rightarrow \quad \frac{Q_3 - \mu}{\sigma} &= 0,675 && \text{Percentil de } Z \sim N(0,1) \\ \Rightarrow \quad Q_3 &= \mu + 0,675\sigma \end{aligned}$$

A partir de estas expresiones se obtiene que el rango intercuartílico está dado por:

$$RI = Q_3 - Q_1 = 1,35\sigma. \tag{2}$$

Note que RI mide la dispersión central de los datos. Tukey et al. (1977) eligió de manera empírica multiplicar el rango intercuartílico por $1,5$, de tal manera que la mayoría delos datos, bajo distribución normal, queden dentro de los bigotes y solo unos pocos sean identificados como atípicos (outiliers). Luego, sustituyendo el resultado final de la ecuación (1) y (2) en la expresión relativa a la BI_1 se obtiene:

$$\begin{aligned} BI_1 &= Q_1 - 1,5 \times 1,35\sigma \\ &= \mu - 0,675\sigma - 1,5 \times 1,35\sigma \\ &= \mu - 2,75\sigma. \end{aligned}$$

Calculando la probabilidad respectiva se tiene:

$$\begin{aligned} P(X \leq BI_1) &= P(X \leq \mu - 2,75\sigma) \\ &= P\left(Z \leq \frac{\mu - 2,75\sigma - \mu}{\sigma}\right) \\ &= P(Z \leq -2,7) \\ &= 0,0035 \end{aligned}$$

Con lo cual se puede afirmar que el $0,7\,\%$ de los datos son datos atípicos ($0,35\,\%$ de cada lado por simetría de la distribución normal). Análogamente se obtiene BI_2. Tukey et al. (1977) repitió el análisis empírico anterior para determinar las barreras externas, para ello estableció multiplicar por 3 a RI, con ello los datos fuera de estos límites representan aproximadamente el $0,0002\,\%$ ($0,0001\,\%$ para cada lado).
Luego, de acuerdo con las barreras, el bigote correspondiente al lado izquierdo será el dato observado mayor o igual a BI_1, análogamente, el bigote del lado derecho será el valor observado menor

o igual a BI_2. Es necesario indicar que la igualdad entre el dato mínimo y el bigote inferior, y/o el dato máximo con el bigote superior, ocurrirá cuando no existan datos atípicos.
Es importante señalar que las barreras externas entregan información respecto a si un dato en particular es considerado atípico o extremo, es decir, si una vez dibujados los bigotes hay observaciones que quedan fuera de ellos diremos que: a) si el dato en cuestión está ubicado entre el bigote y la barrera externa, entonces esta observación es atípica, b) si el dato está fuera de las barreras externas, entonces el dato es clasificado como extremo. Ante cualquiera de estos casos es recomendable verificar si la observación es real o se trata de un error, por ejemplo de tipeo o medición. En caso que se pueda comprobar algún error, esta observación puede ser extraída del conjunto de datos y se debe repetir el análisis.

La Figura 14 resume los pasos necesarios para construir el boxplot o gráfico de cajas, de acuerdo a lo expuesto previamente.

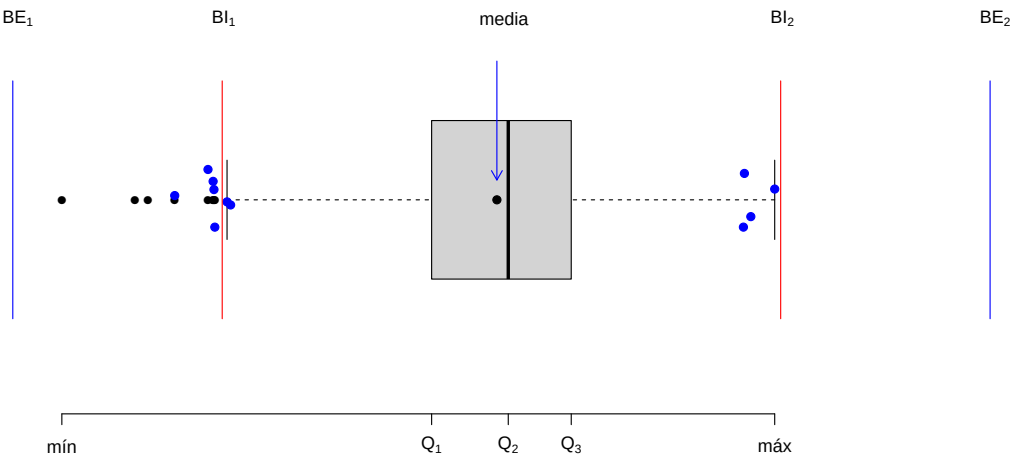


Figura 14: Resumen de los pasos para construir un boxplot.

Destaquemos algunos detalles importantes presentes en el gráfico:

- Los puntos azules corresponden a las observaciones cercanas a las barreras, fueron dibujadas a propósito para ilustrar el proceso de dibujar los bigotes, y de manera que no coincidan con los valores ubicados en la horizontal. Entonces observe que las líneas rojas verticales son los valores de las barreras internas. En particular, al lado izquierdo se observa que hay datos a cada lado de BI_1, pero recordemos que las barreras son imaginarias y no necesariamente corresponden a un dato observado, por ende el bigote se dibuja hasta el valor inmediatamente mayor a a dicha barrera. Análogamente, se observa que al lado derecho ningún valor es mayor a BI_2, por lo cual el bigote derecho se dibujó hasta el valor máximo.
- Para diferenciar entre valores atípicos y extremos se obtienen las barreras externas, en la figura representadas con líneas verticales de color azul. En el lado izquierdo del gráfico se observa que todas las observaciones (puntos negros) están contenidos entre las barreras, por ende este conjunto de datos no posee datos extremos, sólo atípicos. En particular la barrera externa del lado derecho sólo se dibujó por ejercicio, pues como no quedaron valores fuera de la barrera interna derecha no era necesario ubicar a BE_2.
- Al analizar las posiciones relativas, podemos ver que la primera porción de la caja (entre el cuartil 1 y 2) es más grande que la segunda (entre el cuartil 2 y 3), que la media no coincide con la mediana ($\bar{x} < Me$), y que el bigote izquierdo es levemente más largo que el bigote derecho, diferencia que

se vería incrementada si dibujaramos los bigotes hasta el mínimo. Todas estas características nos dan indicios de que la distribución de los datos presenta asimetría negativa.

Consideremos el histograma de esta situación, y nuevamente el boxplot, pero esta vez con la escala de valores de la variable respectiva.

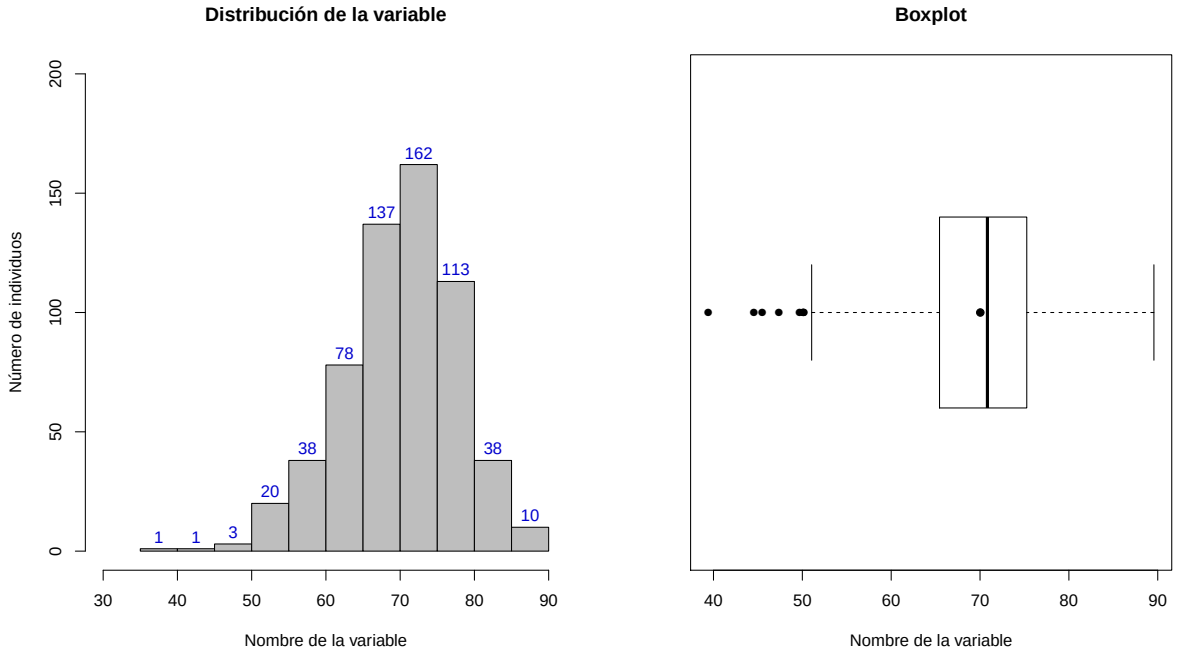


Figura 15: Histograma y Boxplot de una variable cualquiera.

En el histograma mostrado en la parte izquierda de la Figura 15 se observa una asimetría negativa, ya que la cola izquierda es más larga que la derecha. Además, la mayor parte de la distribución de la variable se concentra en los valores más altos. Estas observaciones coinciden con las conclusiones obtenidas previamente a partir del gráfico de caja (boxplot), que también permitió identificar la presencia de datos atípicos, los cuales se evidencian en el histograma como aquellas clases con menor frecuencia absoluta ($f_i = 1$).

3.2.2. Didáctico

Representaciones.

En el aprendizaje y enseñanza de la estadística y las probabilidades se usan representaciones de objetos conceptuales y/o procedimentales. Más aún, las conexiones entre ellas entregan oportunidades a los estudiantes para generar nuevos conocimientos o enriquecer los ya establecidos en su dominio cognitivo. Ciertamente la exposición de los estudiantes a gráficos y tablas, así como también a diversos conceptos estadísticos conceptuales o procedimental provoca en ellos un proceso de interpretación y decodificación, lo que comprende el uso de un lenguaje de representaciones estadísticas. La literatura reporta que cuando los estudiantes aprenden a representar, analizar y hacer conexiones entre ideas matemáticas de múltiples formas, demuestran un entendimiento más profundo, así como el progreso de sus habilidades para resolver problemas Fuson et al. (2005). Esto, debido a que en vista de la naturaleza abstracta de las matemáticas y la estadística, las personas tienen acceso a las ideas matemáticas y estadísticas sólo mediante representaciones.

Lesh et al. (1987) plantean 5 tipos de representación que se conectan para establecer conceptos matemáticos, y en este caso estadísticos (ver Figura 16).

A continuación, se presenta una pequeña descripción de cada tipo de representación:

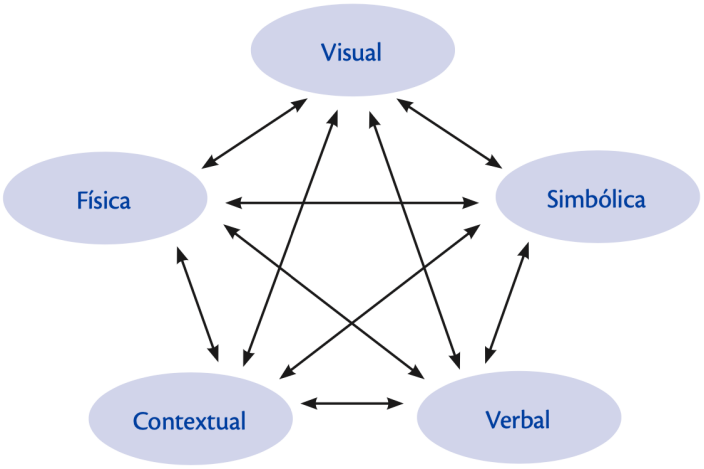


Figura 16: Tipos de representaciones

- Contextual: Las representaciones contextuales se refieren a “guiones” que los estudiantes han experimentado, en los cuales el conocimiento se organiza alrededor de eventos del mundo real. Estos contexto sirven para interpretar y resolver problemas.
- Concreta: Las representaciones concretas se refieren a modelos manipulables como bloques multibase, barras fraccionarias entre otros recursos, los cuales tienen por si solos algún tipo de significado en el conocimiento matemático.
- Visual: Las representaciones visuales se refieren a los diagramas o modelos figurativos de las representaciones concretas, los cuales pueden ser internalizadas como imágenes de los conocimientos matemáticos.
- Verbales: Las representaciones verbales se refieren a la comunicación entorno a la actividad usando la lengua materna de los estudiantes de manera no simbólica o usando un lenguaje matemático.
- Simbólicas: Las representaciones simbólicas se refieren al uso de lenguaje matemático específico.

Dada la clasificación, los estudiantes pueden tener acciones y propósitos diferentes para cada uno de ellos. En cuanto a las acciones, estas pueden ser: a)usar, b) crear o bien, c) modificar. Luego, con la acción establecida, se pueden tener los siguientes propósitos: explicar, argumentar, conectar, interpretar, construir o describir ideas o conceptos estadísticos.

Conexiones entre múltiples representaciones externas (MER por sus siglas en inglés)

Tal como se ha mencionado, la comprensión de los estudiantes se profundiza mediante el análisis de las similitudes existentes en aquellas representaciones que revelan estructuras matemáticas subyacentes o características esenciales de las ideas matemáticas que persisten, independientemente de la forma. Por tal motivo Ainsworth (1999) plantea tres tipos de conexiones (ver Figura 17):

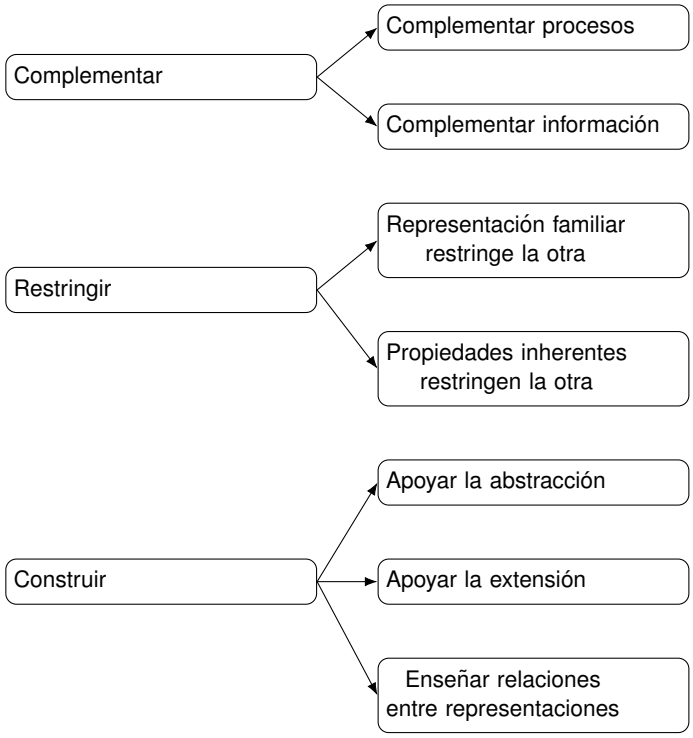


Figura 17. Taxonomía funcional de las representaciones múltiples

- Complementario: Una razón para explotar los MER en entornos de aprendizaje es aprovechar las representaciones que tienen roles complementarios, donde las diferencias entre las representaciones pueden estar en la información que cada una aporta o en los procesos que cada una respalda. Al combinar representaciones que se complementan entre sí de esta manera, se prevé que los alumnos se beneficien de la suma de sus ventajas.
- Constraída: Un segundo uso de las representaciones múltiples es ayudar a los alumnos a desarrollar una mejor comprensión de un dominio mediante el uso de una representación para restringir su interpretación de una segunda representación. Esto se puede lograr de dos maneras: empleando una representación familiar para respaldar la interpretación de una menos familiar o más abstracta, o explotando las propiedades inherentes de una representación para restringir la interpretación de una segunda.
- Profunda: Se ha afirmado que la exposición a múltiples representaciones conduce a una comprensión más profunda. Por ejemplo, Kaput (1989) propone que “el enlace cognitivo de las representaciones crea un todo que es más que la suma de sus partes. Nos permite “ver ideas complejas de una manera nueva y aplicarlas de manera más efectiva”. Así, en este documeno, se considerará la “comprensión más profunda“ en términos del uso de MER para promover la abstracción, fomentar la generalización y enseñar la relación entre las representaciones.

3.3. Orientaciones desde la implementación

A continuación se presenta una serie de recomendaciones para la implementación del problema, las cuales han surgido desde la experiencia de la aplicación de este en el aula.

1. Activación de conocimientos previos. Antes de presentar el problema, se recomienda comenzar una conversación guiada, que considere preguntas y/o actividades tales como:

    a) ¿Recuerdan los elementos que permiten describir un conjunto de datos? Esta pregunta hace alusión a los elementos que caracterizan una distribución, tales como: rango, media, mediana, cuartiles, dispersión y posibles valores atípicos. Esto permitirá que comprendan mejor qué aspectos del histograma deberán vincular luego con el boxplot.

b) Explorar interpretaciones de histogramas conocidos, se recomienda presentar histogramas simples y pedir a los estudiantes que describan verbalmente lo que observan:
   - ¿Dónde se concentran los datos?
   - ¿Hay simetría o asimetría?
   - ¿Observas datos atípicos? Si no ha introducido este concepto defínalos como un valor que se aleja del patrón general de un conjunto de datos. Es necesario indicar que el histograma no indica directamente los outliers, pero sí permite sospechar su existencia al observar por ejemplo: barras aisladas del resto de la distribución, distribución con colas largas o asimétricas (si la cola derecha o izquierda se extiende con pocas observaciones, frecuencias muy bajas en los extremos).

2. Fomentar la conexión entre representaciones. Antes de construir el boxplot, se recomienda que los estudiantes planteen conjeturas sobre su posible forma a partir de la observación del histograma. Esto favorece el desarrollo del pensamiento estadístico, al incentivar la anticipación, la estimación y la validación con base en evidencias gráficas. Para ello se puede apoyar en los elementos estructurales del histograma que pueden ser usados para construir el boxplot (mediana, cuartiles, presencia de posibles datos atípicos)

3. Incorporar la comparación y la reflexión. Una vez elaborado el boxplot, se recomienda contrastar la representación obtenida con el histograma original, analizando semejanzas, diferencias y la información que cada uno permite destacar. Este proceso debe conducir a reflexiones sobre la utilidad de cada tipo de gráfico y las decisiones que puede apoyar en contextos reales.

4. Poner atención a posibles dificultades conceptuales. Es frecuente que los estudiantes interpreten el boxplot como una representación de frecuencias absolutas o que no logren vincular los cuartiles con las áreas del histograma, por lo cual se recomienda reforzar la definición de estos.

4. Problema 3

4.1. Planificación del Problema

a) **Identificación curricular y de contenido del problema**

Objetivo de Aprendizaje del Currículo Escolar	Objetivo de Aprendizaje de la Actividad	Contenidos Estadísticos y de Probabilidad
OA1. Argumentar y comunicar decisiones a partir del análisis crítico de información presente en histogramas, polígonos de frecuencia, frecuencia acumulada, diagramas de cajón y nube de puntos, incluyendo el uso de herramientas digitales.	Establecer relaciones entre diferentes gráficos, extrayendo e interpretando información de uno para ser representada en otro.	■ Cuartiles ■ Boxplot ■ Polígono suavizado

b) **Problema** Suavizando hacia la caja.

El gráfico de distribución suavizado muestra el número de gigabytes que gasta un grupo de estudiantes en un mes de celular.

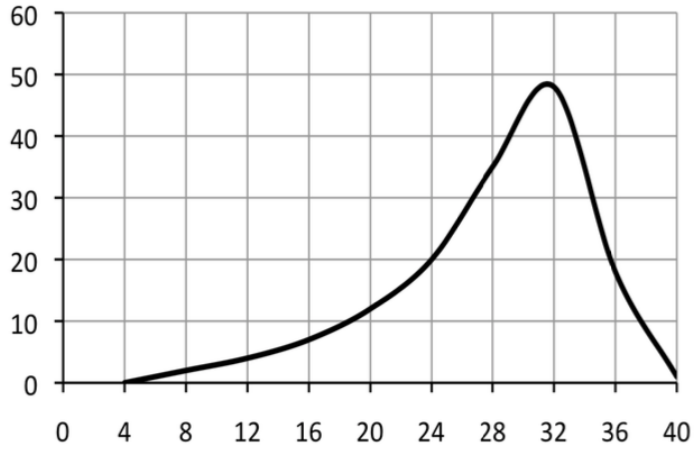
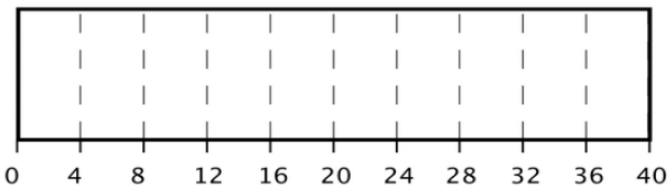


Gráfico de caja



Construye el posible gráfico de caja asociado.

c) **Una solución del problema**

La Figura 17, muestra una solución al problema

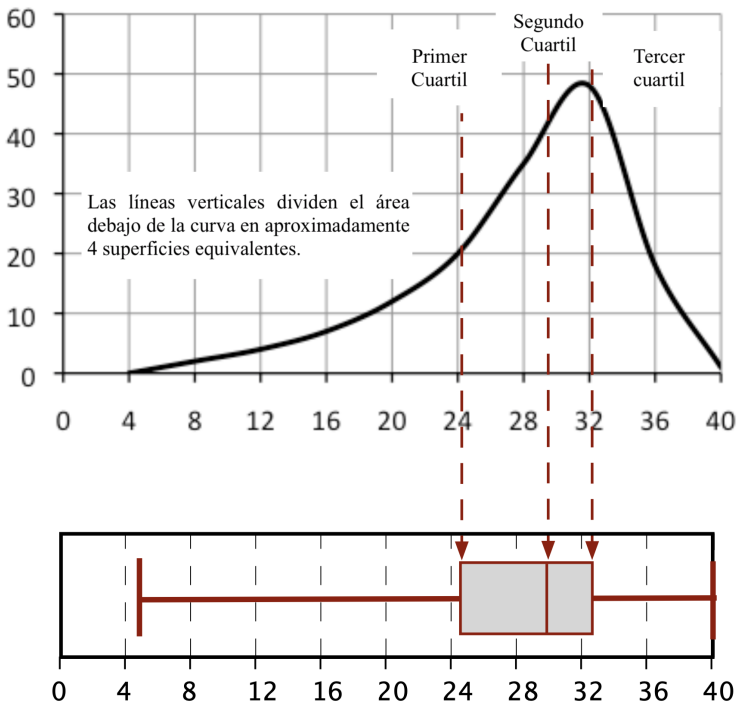
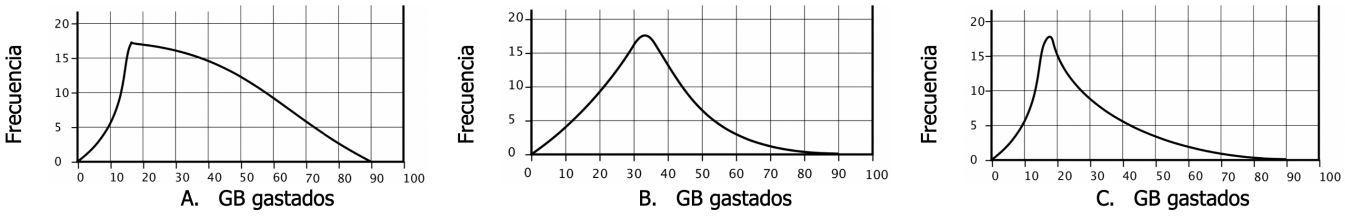


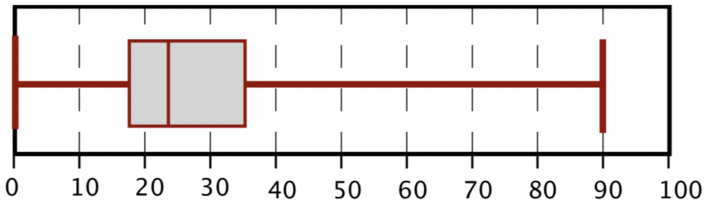
Figura 17: Una posible solución

La mediana debe ser menor que la moda (app. 31). La mediana debe estar ligeramente más cerca al tercer cuartil que del primer cuartil. A modo de explicación, los estudiantes deben dividir el área bajo el gráfico de frecuencia en cuatro áreas iguales, marcando los cuartiles inferior y superior y mediana.

d) **Una simplificación**

Los gráficos de distribución suavizado muestran el número de gigabyte que gasta un grupo de estudiantes en un mes de celular. ¿Cuál de los gráficos corresponde al gráfico de caja?





e) **Formas de consolidar el conocimiento o habilidades pretendidas (preguntas)**

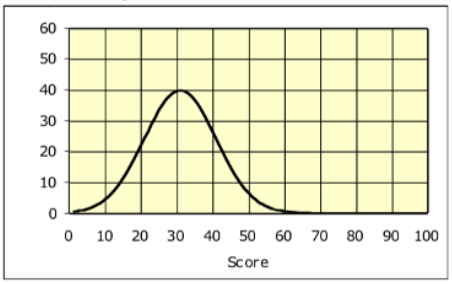
Para consolidar el conocimiento realice las siguientes preguntas cuándo el grupo diga que tiene alguna respuesta a la situación:

- ¿Qué información extrajeron del polígono suavizado para determinar los cuartiles del diagrama de caja?
- ¿Cómo determinaron esa información en el polígono suavizado?
- ¿Existen valores atípicos o fuera de rango? ¿Por qué?
- ¿Son absolutos los valores de cuartíles determinados?
- ¿Qué relación existe entre el RIC y el polígono suavizado?

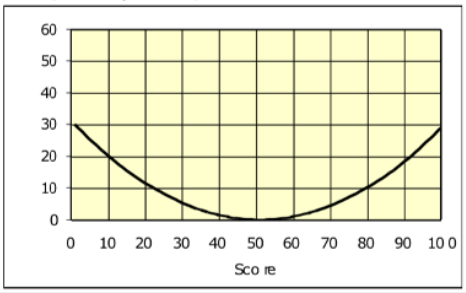
f) **Una extensión**

Relaciona cada distribución suave con el gráfico de caja correspondiente

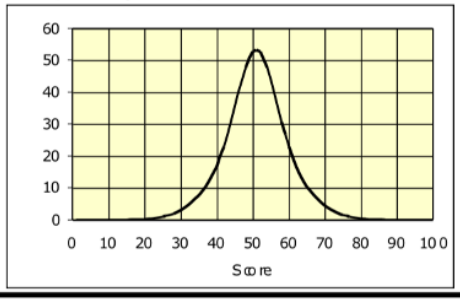
Frequency Graph A



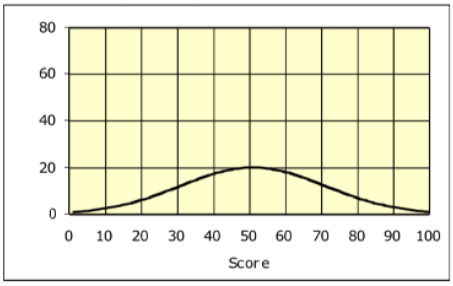
Frequency Graph B



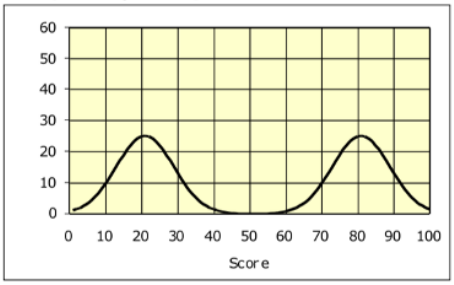
Frequency Graph C



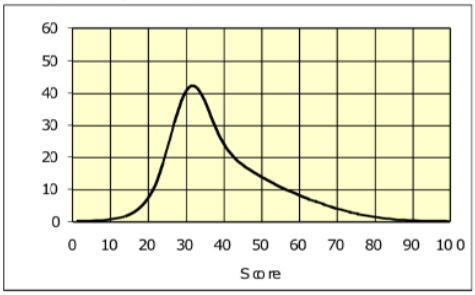
Frequency Graph D



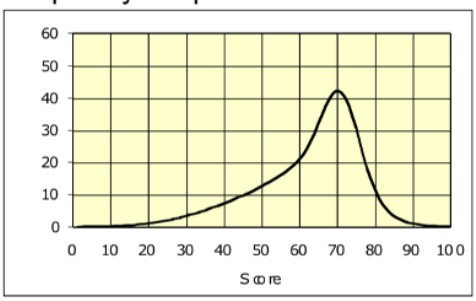
Frequency Graph E



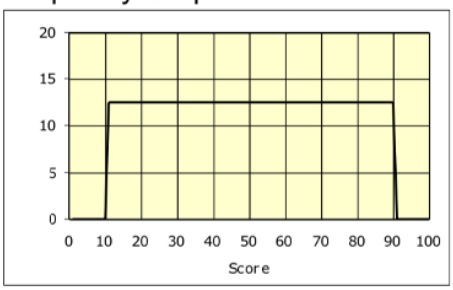
Frequency Graph F

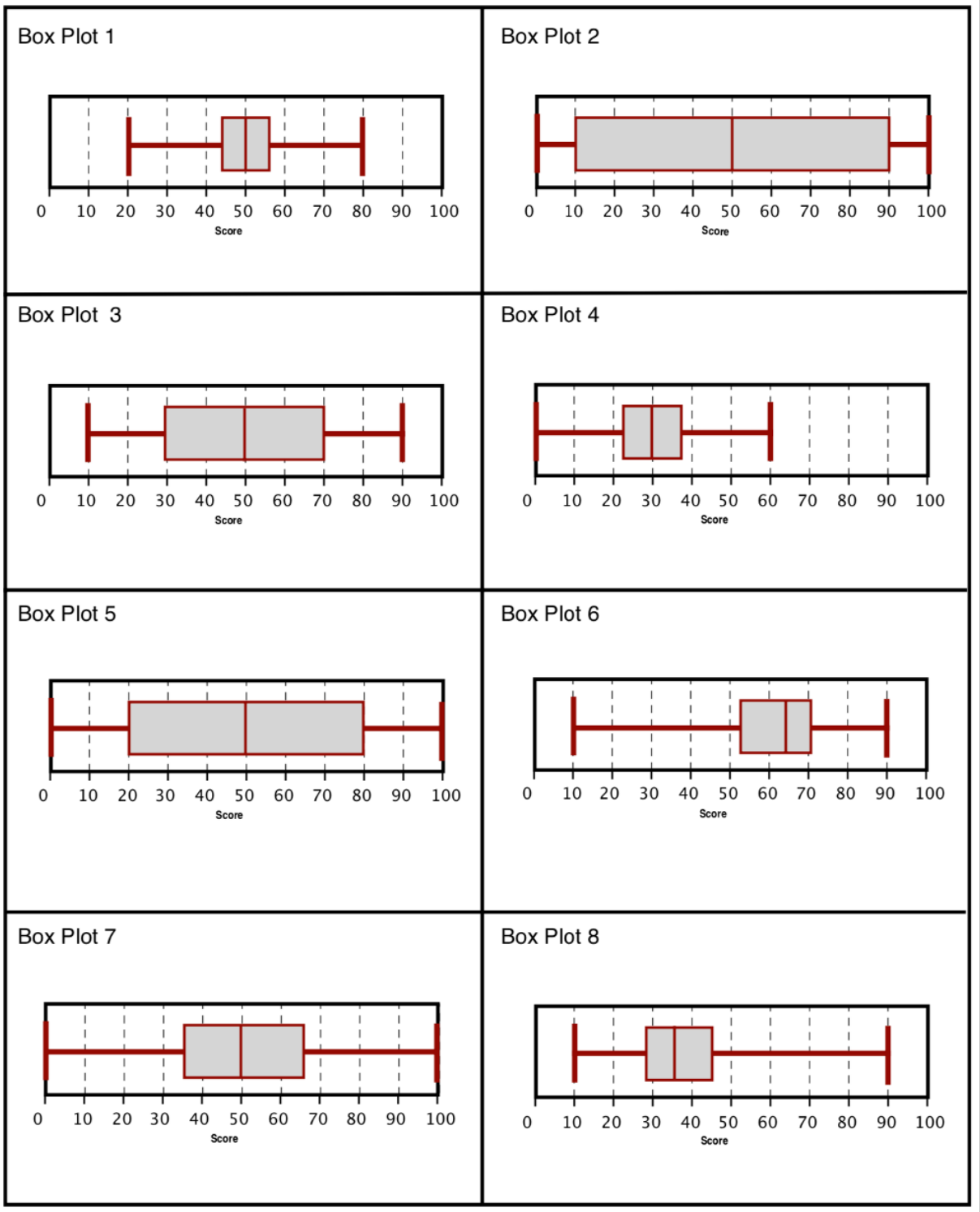


Frequency Graph G



Frequency Graph H





g) **Plenaria**

Para la plenaria se deben seleccionar respuestas que faciliten la discusión entre los estudiantes respecto de: a) elección de información extraída del polígono suavizado para construir el diagrama de caja; b) extrategias de extracción de la información y construcción del polígono. Las respuestas que pueden ser elegidas deben considerar las siguientes características:

- Considerar el gráfico suavizado como una distribución de datos agrupados, por lo que se buscan representantes del agrupamiento para establecer los cuartiles. Considerar el área bajo la curva para distribuirla en cuatro superficies equivalentes.

- Establecer representantes de clases para cada intervalo como la marca de clase; comparar áreas, usando los cuadros que la grilla del gráfico suavizado entrega.

4.2. Contenido para el profesor.

4.2.1. Disciplinar

El polígono de frecuencia es una representación gráfica, la cual consiste en unir por medio de segmentos lineales los puntos formados por los valores de la variable (cuando sólo se consideran clases individuales) o la marca de clases (cuando se considera la construcción de intervalos) y su respectiva frecuencia. De lo anterior, es claro que un polígono de frecuencia sólo se puede elaborar cuando la variable analizada es de tipo cuantitativa. La Figura 18, muestra un ejemplo en ambos casos:

- La Figura 18a) corresponde al polígono de frecuencias de los datos de la Tabla 7, donde recordemos que la variable en cuestión es cuantitativa continua y que a partir de la información disponible se construyó una tabla de frecuencias considerando clases individuales. El gráfico en cuestión está sobrepuesto sobre el gráfico de barras correspondiente a la situación de interés, para ello se dibujaron puntos considerando el punto central de cada barra y la respectiva frecuencia absoluta, y luego estos puntos fueron unidos por segmentos de línea, adicionalmente y con el propósito de que efectivamente sea un polígono, por ende cerrado, se añaden al gráfico dos puntos, el primero de ellos antes del primer valor de la variable, el cual claramente tendrá frecuencia cero, pues no es un valor observado de la variable, y el segundo después del último valor observado de la variable (también con frecuencia nula).
- Mientras que la Figura 18b) muestra el polígono de frecuencias sobre un histograma de una variable hipotética, lo cual implica la construcción previa de intervalos. La única diferencia con el gráfico a), es que los puntos son construidos con la marca de clase de cada intervalo, sin embargo, no olvidemos que estas también son el punto medio de cada barra. En este caso para cerrar el polígono, se añaden dos intervalos de la misma amplitud que los previamente construidos, uno antes del primer intervalo, y otro después del último, ambos con frecuencia nula.

Es importante hacer notar que la un polígono de frecuencias puede ser construido con la frecuencia absoluta, la frecuencia relativa o la frecuencia relativa porcentual.

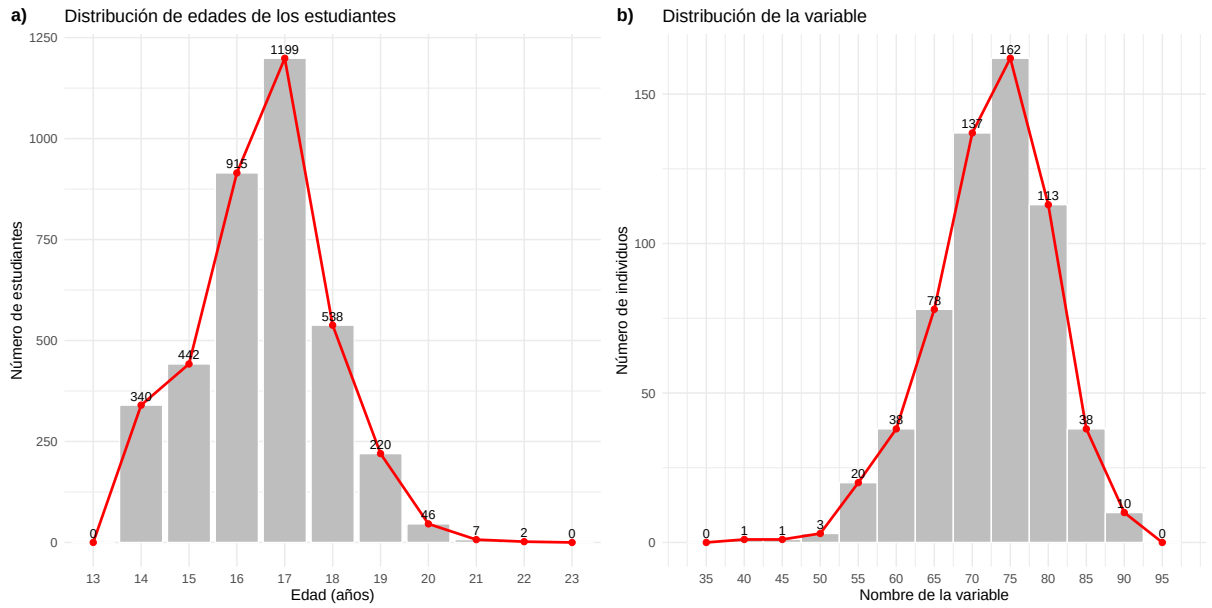


Figura 18: Polígono de frecuencia datos agrupados en clases individuales a) y por intervalos b).

Básicamente la utilidad de los polígonos de frecuencia es que permiten analizar de manera gráfica la distribución de los datos, concluir sobre la presencia de simetría o de asimetría en ellos, donde esta última de acuerdo con la Figura 12 puede ser positiva o negativa. Para lograr este objetivo es usual suavizar el polígono de frecuencias, lo cual simplemente consiste en trazar una curva a mano alzada por sobre el polígono construido. De acuerdo con esto, en la Figura 19 se muestran los polígonos de frecuencia suavizados en cada una de las situaciones anteriores, las cuales están dibujadas de color azul.

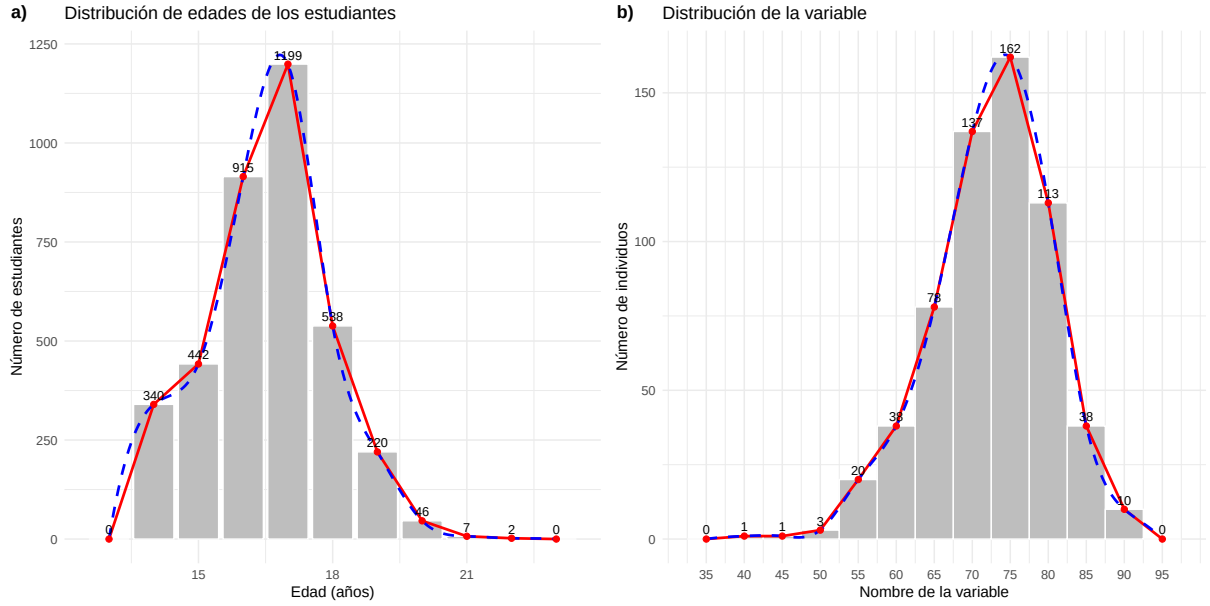


Figura 19: Polígono de frecuencia suavizados para datos agrupados en clases individuales a) y por intervalos b).

El polígono de frecuencias está estrechamente relacionado con otras representaciones gráficas de la variable de interés, como los gráficos de barras o los histogramas. De manera análoga, también se relaciona con el boxplot, el cual no solo resulta útil para analizar la distribución de los datos, sino que además permite identificar la presencia de datos atípicos.

En los ejemplos anteriores se observa que, en el polígono de frecuencias a) no es posible determinar visualmente la asimetría de los datos; sin embargo, se identifican posibles valores atípicos en los niveles más altos de la variable (20, 21 y 22 años), debido a su baja frecuencia. En contraste, en el polígono de frecuencias b) se aprecia una asimetría negativa, junto con la presencia de posibles valores atípicos en los niveles más bajos de la variable.

Este análisis puede complementarse con el respectivo boxplot. La Figura 20 muestra, para cada situación planteada, el gráfico de barra o histograma , el polígono de frecuencia suavizado y el boxplot correspondiente. En el caso a), el boxplot permite concluir que existen valores atípicos en 14, 19, 20, 21 y 22 años; que el 50 % de los estudiantes tienen entre 15 y 16 años; que el tercer cuartil coincide con la mediana, y que el promedio es menor que esta. En el caso b), el boxplot revela la existencia de valores atípicos en los niveles más bajos de la variable (menores a 50 aproximadamente), que el promedio es ligeramente inferior a la mediana, y que la proporción izquierda de la caja es levemente mayor que la derecha (la mediana se encuentra ligeramente más próxima al tercer cuartil). Estas características, sumadas al hecho de que el polígono de frecuencias es unimodal, permiten concluir que los datos presentan efectivamente una asimetría negativa, conclusión que no es posible establecer en el ejemplo a).

4.2.2. Didáctico

Transnumeración

Los polígonos suavizados y los diagramas de cajas, así como los gráficos de barras y otros, son una transformación de datos sin procesar a una forma completamente diferente. La comprensión de esta transformación

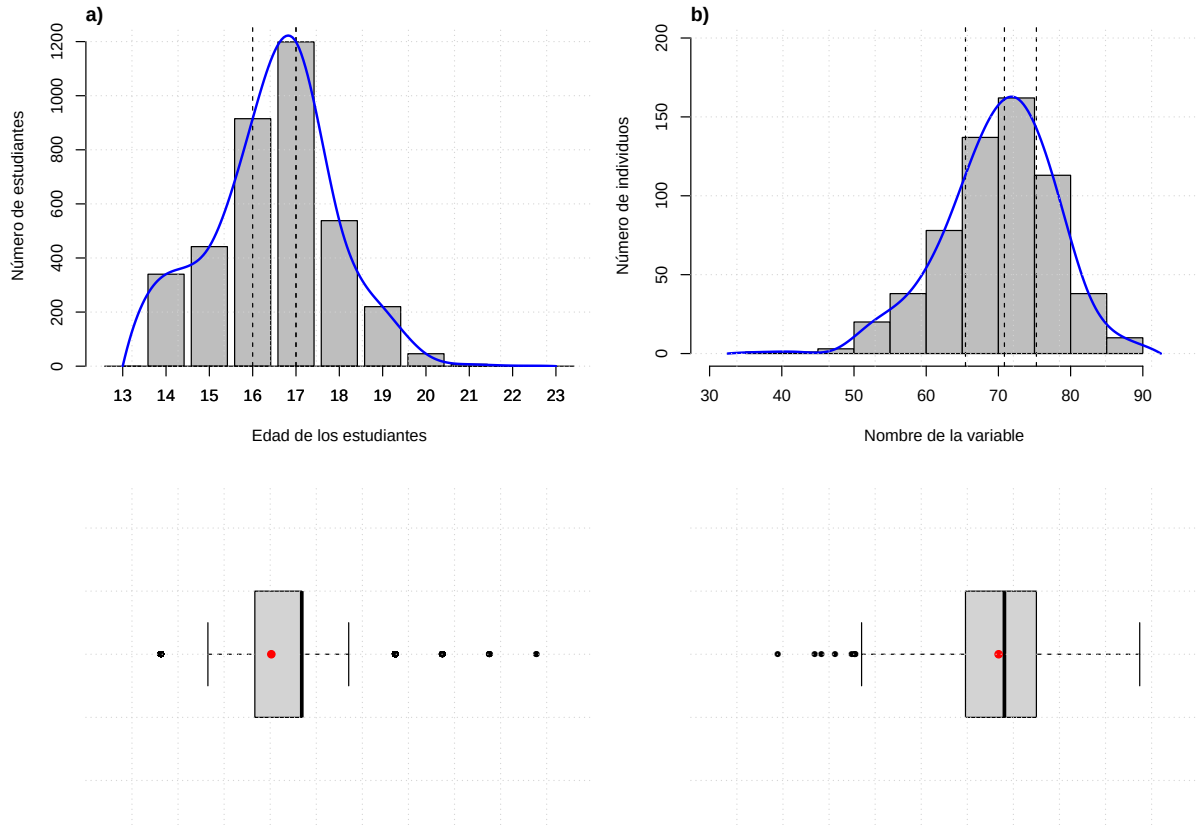


Figura 20: Polígono de frecuencia suavizados y boxplot para datos agrupados en clases individuales y por intervalos.

es un desafío, y la enseñanza de la estadística debe encontrar formas de respaldarla. Tal transformación cambia la representación de los datos, un proceso que Wild y Pfannkuch (1999a) han definido como ***transumeración***, y es uno de los marcos fundamentales para el razonamiento estadístico, para una mejor comprensión de la variación, la distribución y muchos otros conceptos estadísticos importantes.

Una de las ideas fundamentales del aprendizaje de la estadística es la de formar y cambiar representaciones de datos de un sistema para llegar a una mejor comprensión de este. Así, se asigna la palabra transnumeración a esta idea. Lo definimos como ”transformaciones numéricas hechas para facilitar la comprensión”. La transnumeración ocurre cuando encontramos formas de obtener datos (a través de la medición o clasificación) que capturan elementos significativos del sistema original. Impregna todo el análisis de datos estadísticos y ocurre cada vez que cambiamos nuestra forma de ver los datos con la esperanza de que esto nos transmita un nuevo significado. Podemos mirar a través de muchas representaciones gráficas para encontrar algunas realmente informativas. Podemos volver a expresar los datos a través de transformaciones y reclasificaciones en busca de nuevos conocimientos. Podríamos probar una variedad de modelos estadísticos. Y al final del proceso, la transnumeración ocurre una vez más cuando descubrimos representaciones de datos que ayudan a transmitir nuestra nueva comprensión sobre el sistema real a los demás. La transnumeración es un proceso dinámico de cambio de representaciones para generar comprensión.

Comparar distribuciones

La tarea de comparar en el aprendizaje de estadística es tan básica como lectura de formas de representar y organizar. Pfannkuch (2006) plantea una serie de focos al realizar comparaciones, a saber:

a) Generación de hipótesis. Compara y razona sobre la tendencia del grupo.

b) Resumen. Compara puntos de resumen de números equivalentes y no equivalentes.

c) Cambio: Compara un diagrama de caja con otro y se refiere al cambio.

d) Señal: Compara la superposición del 50 % central de los datos.

e) Propagación. Compara y se refiere al tipo de dispersión/densidades a nivel local y global dentro y entre diagramas de caja.

f) Muestreo. Considera el tamaño de la muestra, la comparación si se tomará otra muestra, la población sobre la cual hacer una inferencia.

g) Explicativo. Comprende el contexto de los datos, considera si los hallazgos tienen sentido, considera explicaciones alternativas para los hallazgos.

h) Caso individual. Considera posibles valores atípicos, compara casos individuales.

i) Evaluativo. Evidencia descrita, evaluada en su solidez, sopesada.

j) Referente. Etiqueta de grupo, medida de datos, medida estadística, atribución de datos, distribución de gráficos de datos, conocimiento contextual y estadístico

En esta línea, Makar y Confrey (2004) plantean niveles de comparación, los cuales se describen a continuación.

a) En un nivel ***Pre-descriptivo***, no se hace ningún reconocimiento de relaciones entre conjuntos de datos, excepto en base a puntos de datos individuales o evidencia anecdótica. Si se hacen conjeturas a este nivel, son inconmensurables.

b) Los estudiantes que usan un nivel ***Descriptivo*** se enfocan en estadísticas de resumen y hacen comparaciones absolutas entre conjuntos de datos sin tener en cuenta la variabilidad. Las conjeturas asumen que los datos están infinitamente disponibles para responder cualquier pregunta.

c) La primera vista holística de los datos ocurre en el nivel de ***distribución emergente***, donde se utilizan descriptores cualitativos informales de los datos, junto con estadísticas de resumen básicos, para describir dos conjuntos de datos. Los estudiantes comienzan a comprender la dificultad de crear conjeturas medibles, pero no pueden resolver con éxito el conflicto y muestran frustración al intentar escribir una conjetura adecuada. La variabilidad, aunque reconocida, no se entiende más allá de un nivel descriptivo.

d) Los estudiantes con una ***Visión Transicional*** de los datos comienzan a comprender la influencia de la variabilidad al comparar dos grupos. Se muestra más flexibilidad (por ejemplo, representaciones gráficas múltiples, medidas alternativas de centro o dispersión) al comparar conjuntos de datos en este nivel. Las conjeturas, aunque cuestionablemente medibles, han progresado para mostrar una comprensión elemental de la dificultad de crear una conjetura que no comprometa demasiado la pregunta en cuestión, pero que permita la posible recopilación de datos. El concepto de tendencia estadística pasa a formar parte de la discusión y conclusión sobre los datos.

e) Finalmente, en el nivel de ***Estadística Emergente***, los estudiantes ganan confianza en el uso de estadísticas descriptivas estándar para comparar conjuntos de datos, teniendo en cuenta las diferencias entre las medidas del centro a la luz de la variabilidad en los datos y los tamaños de muestra de los conjuntos de datos. Las conjeturas demuestran cierta capacidad para formular preguntas que equilibren las restricciones de datos con el problema en cuestión. El contexto y las descripciones cuantificadas están bien integrados en las conclusiones y las inferencias pueden intentar basarse en modelos estadísticos, si corresponde.

4.3. Orientaciones desde la implementación

A continuación se presenta una serie de recomendaciones para la implementación del problema, las cuales han surgido desde la experiencia de la aplicación de este en el aula.

1) Activación de conocimientos previos. Antes de presentar el problema, se recomienda iniciar una conversación guiada que permita movilizar ideas sobre lo que conocen los estudiantes respecto a las representaciones de distribuciones de datos y sus elementos característicos. Algunos ejemplos de preguntas orientadoras son:

a) ¿Qué información nos entrega un polígono de frecuencias sobre un conjunto de datos? Esta pregunta tiene como principal foco que el estudiante conecte la representación solicitada con los elementos que requiere para su elaboración. Además para que identifique características de la representación, tales como, la ubicación aproximada de las medidas de posición, el máximo de la curva (asociada al valor modal), etc.

b) ¿Cómo podríamos describir la forma de una distribución a partir del gráfico? (por ejemplo, si es simétrica, sesgada o presenta concentraciones de datos).

c) ¿Qué relación existe entre el polígono de frecuencias y otras representaciones como el histograma o el boxplot? Esta pregunta ayudará a situar a los estudiantes en el contexto del problema y a reconocer los elementos clave que luego deberán vincular (mediana, cuartiles, dispersión y posibles valores atípicos).

2) Fomentar la interpretación y el análisis visual. Se recomienda que los estudiantes observen cuidadosamente el polígono suavizado y describan lo que perciben:

a) ¿Dónde se concentran los datos?

b) ¿Qué tan extendida está la distribución? Esta pregunta con el propósito que el estudiante concluya sobre la posible asimetría de la distribución.

c) ¿Existen valores que parecen alejados del resto? Este proceso de descripción permite desarrollar la capacidad de interpretar gráficamente patrones de dispersión y tendencia, lo que será esencial para anticipar la forma del boxplot.

3) Promover la conexión entre representaciones. Antes de construir el boxplot, es conveniente que los estudiantes elaboren conjeturas sobre su posible forma a partir del polígono suavizado. Por ejemplo, estimar visualmente dónde podría ubicarse la mediana o los cuartiles, o si se espera presencia de valores atípicos. Esta anticipación favorece la comprensión de cómo distintas representaciones (polígono y boxplot) expresan la misma información de manera complementaria.

4) Comparar y reflexionar sobre las representaciones. Una vez construido el boxplot, se recomienda analizarlo en conjunto con el polígono suavizado, contrastando la información que cada uno resalta:

a) ¿Qué aspectos del conjunto de datos se aprecian mejor en el boxplot que en el polígono?

b) ¿Qué información se pierde o se mantiene al pasar de una representación a otra?

Estas comparaciones permiten fortalecer la comprensión del propósito y la utilidad de cada tipo de gráfico en el análisis de datos.

5) Atender a posibles dificultades conceptuales. Es común que los estudiantes:

- Interpreten el boxplot como una representación de frecuencias en lugar de una de posición y dispersión.
- No logren vincular visualmente los cuartiles con las zonas del polígono suavizado.
- Supongan que los valores atípicos siempre deben aparecer en el polígono, sin comprender su detección formal.

Por ello, se recomienda reforzar las definiciones de mediana, cuartiles y rango intercuartílico, y discutir ejemplos donde los outliers sean o no visibles en la representación suavizada.

5. Problema 4

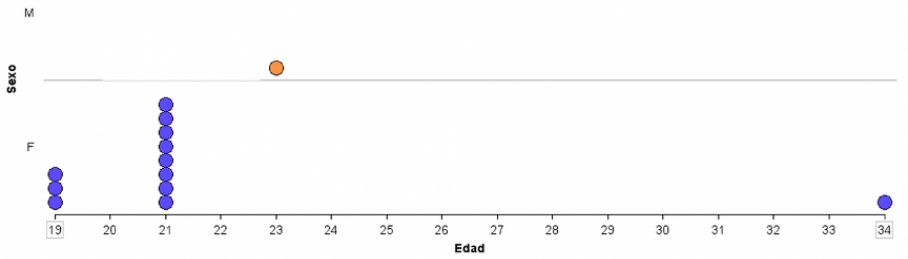
5.1. Planificación del Problema

a) **Identificación curricular y de contenido del problema**

Objetivo de Aprendizaje del Currículo Escolar	Objetivo de Aprendizaje de la Actividad	Contenidos Estadísticos y de Probabilidad
OA 2. Resolver problemas que involucren los conceptos de media muestral, desviación estándar, varianza, coeficiente de variación y correlación muestral entre dos variables, tanto de forma manuscrita como haciendo uso de herramientas tecnológicas digitales.	Analizar la relación entre la distribución de un conjunto de datos y la media como estadístico de resumen.	▪ Media ▪ Variabilidad ▪ Dispersión

b) **Problema**

Los 3 hombres de un curso tienen una edad promedio de 22 años y el curso entero, de 36 estudiantes, tienen una edad promedio de 24 años. Representa los puntos de las personas que faltan en el gráfico de distribución.



c) **Una solución del problema**

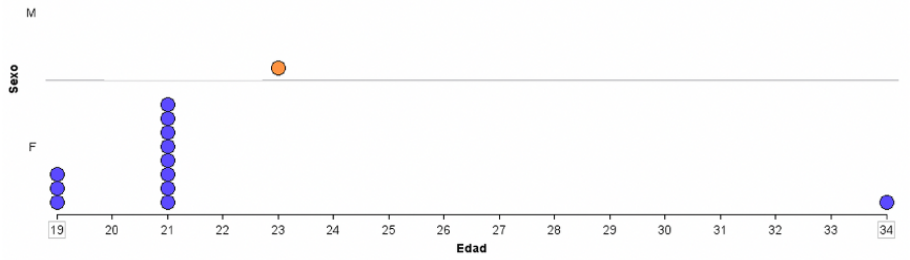
Dibujar las edades en un gráfico de puntos requiere una estrategia que implique comprender la media, tanto a nivel conceptual como procedimental. Para demostrar esta comprensión, se puede comenzar analizando las edades de los hombres como una forma de reducir el problema a un caso más pequeño. Dado que son solo tres hombres con una edad promedio de 22 años, esta información se incorpora directamente en el procedimiento de cálculo de la media.

$$\frac{(E_1 + E_2 + 23)}{3} = 22$$

Si despejamos la suma de las edades

$$E_1 + E_2 = 43$$

Por lo que se deben buscar combinaciones aditivas que tengan como suma 43 años. Restringido a la escala de 19 a 34 años de edad. Las posibles edades para los tres hombres son: $E_1 = 19$ y $E_2 = 24$; $E_1 = 20$ y $E_2 = 23$; $E_1 = 21$ y $E_2 = 22$, sin considerar el orden. Este mismo procedimiento se puede aplicar para obtener las edades de las mujeres del curso.



d) **Una simplificación**

Los 3 hombres de un curso tienen una edad promedio de 22 años y el curso entero, de 36 estudiantes, tienen una edad promedio de 24 años. Representa solo los puntos de los hombres que faltan.

e) **Formas de consolidar el conocimiento o habilidades pretendidas (preguntas)**

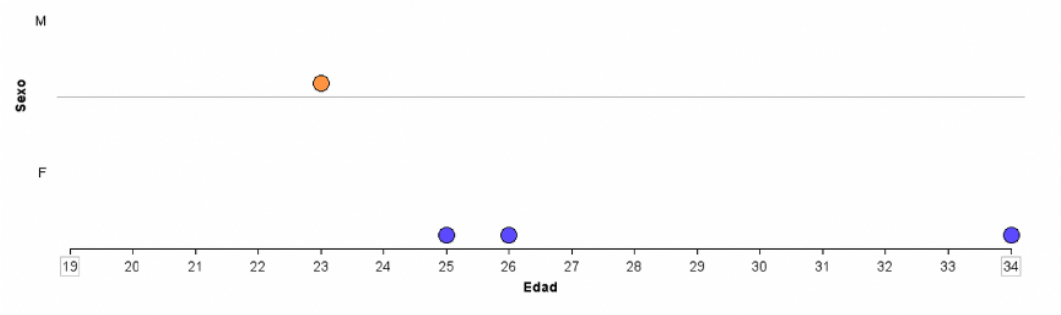
Para consolidar el conocimiento realice las siguientes preguntas cuándo el grupo diga que tiene alguna respuesta a la situación:

- ¿Cuáles son todas las edades de los hombres?
- ¿Por qué son solo esas edades?
- ¿Pueden dar un ejemplo para las edades de las mujeres?
- ¿Cuál sería una estrategia para identificar las edades de las mujeres?
- ¿Cómo se puede extender esta estrategia para conocer todas las edades de las mujeres?

f) **Una extensión**

La Edad Media

La edad promedio de un curso es igual a 24 años. Si se compone de hombres y mujeres. Completa los puntos que faltan en la distribución y determina las edades promedio de cada sexo.



g) **Plenaria**

Para la plenaria se deben seleccionar respuestas que facilite la discusión entre los estudiantes en torno a: a) el conocimiento de las propiedades de la media al establecer estretegias de cálculo; b) la comprensión conceptual de la media, como reparto equitativo de la media; c) el reconocimiento del error de la media total como la media de las medias marginales de subconjuntos de la población total de datos. Las respuestas que pueden ser elegidas deben considerar las siguientes características:

- Representar una aproximación al modelo de equilibrio entre las distancias individuales de las edades a la edad dada.
- Reconocer que existe un sumando que se distribuye entre las diferentes edades de los hombres o bien entre las mujeres.
- Establecer que la media del curso es la media de las medias de los hombres y las mujeres.

5.2. Contenido para el profesor.

5.2.1. Disciplinar

Medidas de tendencia central (MTC).

En capítulos anteriores discutimos cómo la información puede organizarse y representarse mediante tablas de frecuencias y gráficos, siempre considerando la naturaleza de la variable. La utilidad principal de estas herramientas es resumir los datos de manera que sea posible extraer conclusiones sobre la forma de la distribución, identificar concentraciones o dispersión, y detectar la presencia de valores atípicos, entre otros aspectos. Sin embargo, las conclusiones obtenidas únicamente a partir de los gráficos pueden tener un grado de subjetividad. Por ejemplo, la apariencia de un histograma depende de la cantidad y amplitud de los intervalos definidos en su construcción. Analicemos el siguiente ejemplo:

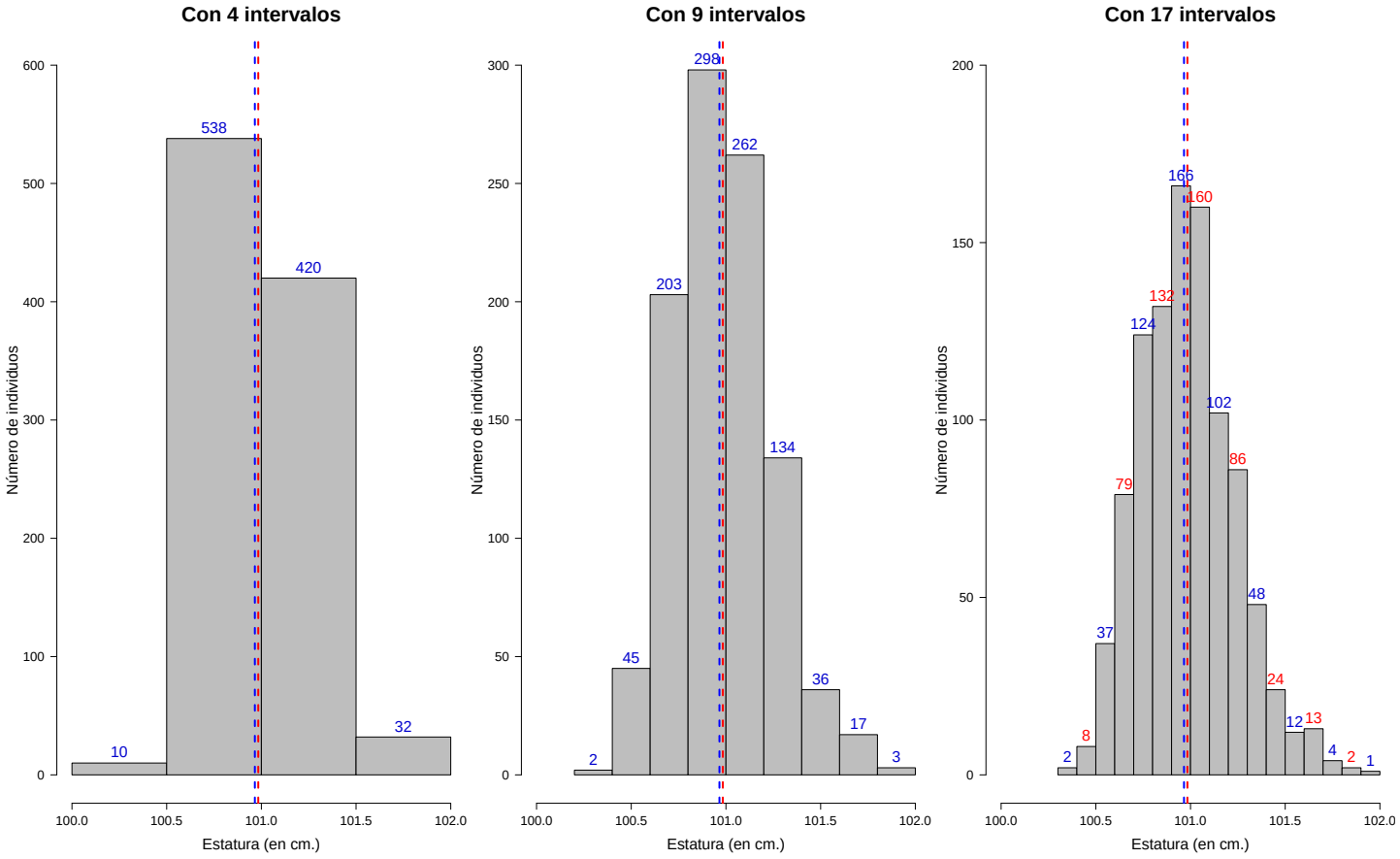


Figura 21: Histogramas con distinto número de intervalos.

En la Figura 21 se muestran distintos histogramas construidos a partir del mismo conjunto de datos (estatura de un grupo de niños en cm). En el primero, elaborado con 4 intervalos, podría concluirse que la mayoría de los niños mide entre 100,5 y 101,0 cm; no obstante, a partir de ese gráfico no es posible determinar la asimetría de la distribución ni identificar valores atípicos. En el segundo histograma, con 9 intervalos, se obtiene una visión más precisa: puede afirmarse que la clase más frecuente corresponde al rango 100,8–101,0 cm, que la distribución presenta asimetría positiva y que probablemente existen valores atípicos. Finalmente, estas conclusiones se hacen aún más evidentes en el tercer gráfico, el cual fue construido con 17 intervalos.

Debido a la subjetividad anterior, es necesario disponer de herramientas que nos permitan cuantificar, de manera tal que sea posible tomar decisiones con respecto a la forma de la distribución, este rol lo cumplen las medidas de tendencia central o simplemente MTC. Note que en las figuras anteriores se

dibujaron con color rojo la media de los datos y en color azul la mediana, las cuales como podemos ver no coinciden, y tal como se indicó en capítulos previos, cuando $\overline{x} > Me$, implica que la distribución de los datos es asimétrica, en particular positiva. Ahora note que si bien estas cantidades parecen numéricamente cercanas (100,9836 cm. versus 100,9656 cm.) debemos considerar que el dato mínimo es 100,3 cm. y el máximo 102,0 cm., lo que hace que el Rango de datos (diferencia entre la observación máxima y mínima) sea pequeño.

Recordemos que la forma de calcular el promedio (media aritmética) depende de cómo estén organizados los datos, sin embargo, es recomendable utilizar los datos sin agrupar para realizar el cálculo respectivo, debido principalmente a que en el caso de datos agrupados por intervalos se requiere la marca de clase, la cual cambiará de acuerdo a la amplitud del intervalo, que a su vez depende del número de intervalos construidos.

Propiedades de la media. Previamente ya señalamos algunas, ahora las formalizaremos matemáticamente:

a) A la diferencia de cada dato al promedio, $x_i - \overline{x}$ se le denomina desviación del $i-$ésimo dato con respecto al promedio. Y se cumple que:

$$\sum_{i=1}^{n}(x_i - \overline{x}) = 0,$$

es decir, la suma de las desviaciones es cero.

b) El promedio de la suma (o diferencia) entre una constante c y la variable X es igual a la suma (o diferencia) de la constante y el promedio de la variable, es decir:

$$\overline{x \pm c} = \overline{x} \pm c,$$

es decir, si a cada valor de la variable se le suma o resta una constante (la misma para todos), el nuevo promedio es igual al promedio anterior más (o menos) dicha constante.

c) El promedio del producto de una constante c por una variable X es igual al producto de esta constante por el promedio de la variable, es decir:

$$\overline{cx} = c\overline{x},$$

en otras palabras, si a cada valor de la variable lo multiplicamos por una constante, entonces el nuevo promedio simplemente se obtiene multiplicando el promedio anterior por esta constante.

d) Si X e Y representan dos variables con el mismo número de datos, entonces el promedio de la suma de estas variables es igual a la suma de los promedios respectivos, es decir:

$$\overline{x + y} = \overline{x} + \overline{y}$$

donde, la cantidad $\overline{x + y}$ representa el promedio global de las dos variables en conjunto. Es importante señalar que para que esta propiedad se cumpla, las variables deben representar lo mismo (expresadas en la misma unidad de medida) y la misma cantidad de datos.

Medidas de variablidad

Debido a que existen conjuntos de datos que tienen la misma media aritmética (promedio) es necesario disponer de herramientas que permitan cuantificar la variabilidad de los datos, con el propósito de seguir describiendo y analizando la distribución de estos. Existen varias opciones, sin embargo la más popular es la desviación estándar, la cual se obtiene a partir de la raíz cuadrada de la varianza, donde el valor de esta última se obtiene a partir de la siguiente expresión:

$$s_X^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1} = \frac{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}{n-1}$$

Note que la varianza se define como el promedio cuadrático de las desviaciones de los datos con respecto al promedio, por ende las unidades de medida de las variables quedan expresadas al cuadrado, razón por la cual, la medida que variabilidad que se interpreta es la desviación estándar. Luego, la desviación estándar representa cuánto se desvían en promedio las observaciones del promedio. Es claro que a mayor varianza mayor desviación estándar, y como consecuencia mayor dispersión de los datos.

Analicemos la varianza en las siguientes situaciones, ver Figura 22, en ella se presentan 4 conjuntos de datos, todos con igual promedio ($\overline{x} = 18$ años), marcada con una línea vertical de color rojo. Note que los dos gráficos superiores tienes $n = 260$ y $n = 130$ datos, respectivamente, donde las frecuencias absolutas del primero son el doble que las del segundo, ¿cuál conjunto cree que presenta mayor varianza y por ende mayor dispersión?. Las dos figuras inferiores tienen el mismo número de datos, ¿cuál diría que tiene mayor dispersión?.

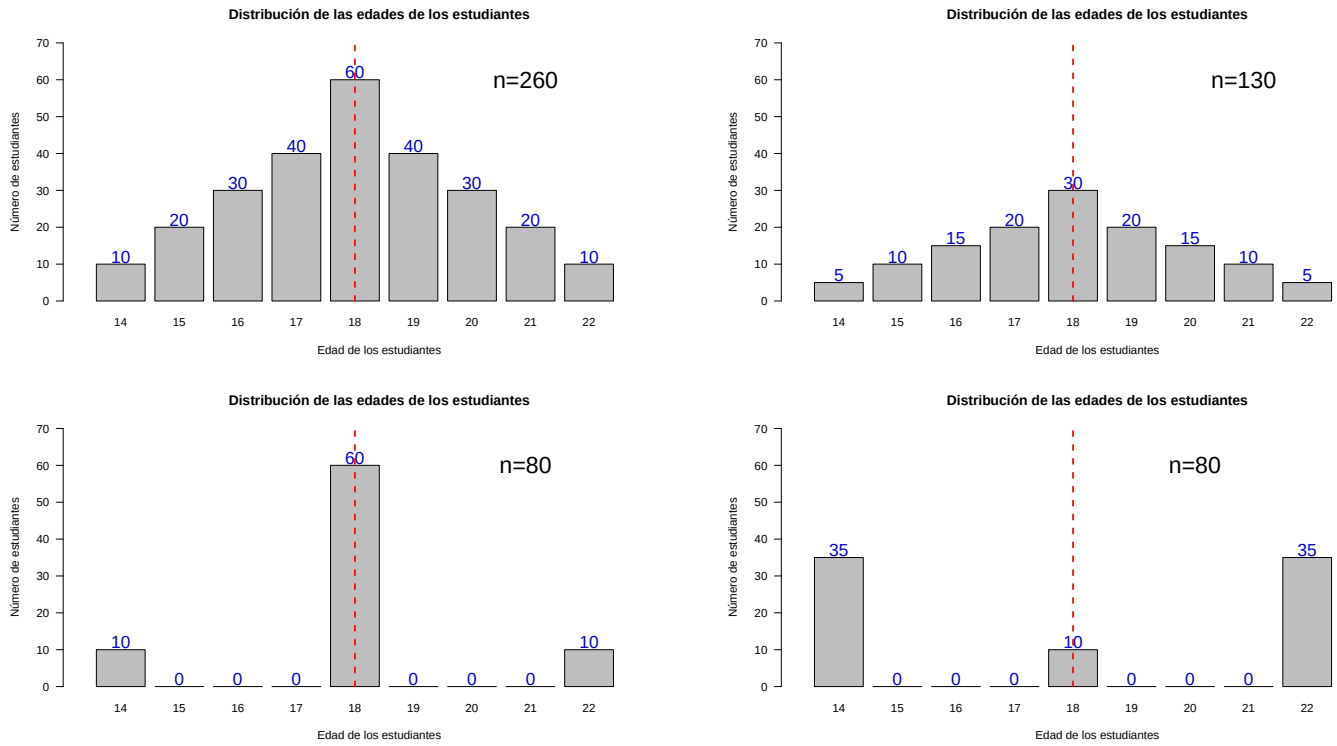


Figura 22: Comparación de varianza con distintos tamaños muestrales.

Para responder a las preguntas anteriores sin requerir cálculos, es necesario observar la forma de la distribución en cada caso, pues si analizamos la fórmula de cálculo se puede concluir que esta básicamente considera el aporte invididual de cada observación, medida como la desviación individual al cuadrado. De acuerdo con lo anterior, podemos observar que el aporte a la expresión de la varianza, en las figuras superiores es aproximadamente igual (caso muestral), debido a que simplemente la primera tiene frecuencias el doble que la segunda. Entonces considerando que la varianza para datos agrupados en intervalos individuales está dada por:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2 \times f_i}{n-1}$$

Se puede establecer una relación entre la varianza del primer conjunto (1) con respecto al segundo (2):

$$s_1^2 = 2\frac{\sum_{i=1}^{n}(x_i - 18)^2 \times f_i}{259} \approx s_2^2,$$

donde f_i es este caso representan las frecuencias absolutas del grupo 2. La aproximación señalada se debe a que la mitad de 259 no es precisamente 130 - 1, numéricamente esto se puede observar en la Tabla 11. Luego, se puede concluir que si dos conjunto de datos tienen la misma distribución, independiente de la cantidad de datos (n), entonces también tiene la misma varianza y/o desviación estándar, por ende, misma dispersión. Por otro lado, si comparamos las figuras inferiores, se puede concluir que a pesar de tener el mismo número de observaciones la segunda de ellas tiene más contribuciones al numerador de la varianza que la primera, esto es debido a que las observaciones extremas tienen mayor frecuencia absoluta. Note que las cuatro figuras se distribuyen de manera simétrica en torno a la media (18 años), sin embargo, solo las superiores tienen forma de campana, propiedad muy importante para la estadística que discutiremos más adelante.

La Tabla 11 muestra las varianzas y desviaciones estándar para los conjuntos de datos Edad en años, representados en la Figura anterior.

Conjunto	1	2	3	4
Varianza (s^2)	3,86	3,88	4,05	14,17
Desviación Estándar (s)	1,96	1,97	2,01	3,77

Tabla 11: Varianza y desviación estándar para los conjuntos de datos Edad en años.

Algunas características y propiedades importantes de la desviación estándar son:

a) $s \geq 0$, la desviación estándar está acotada inferiormente por el 0, sin embargo, no tiene cota superior. En el caso particular cuando $s = 0$, implica que todas las observaciones son iguales al promedio.

b) La desviación estándar de la suma (o diferencia) entre una constante c y la variable X es igual a la desviación estándar de la variable, es decir:

$$s_{X \pm c} = s_X,$$

es decir, si a cada valor de la variable se le suma o resta una constante (la misma para todos), la desviación estándar no se ve afectada por dicha constante.

c) La desviación estándar del producto de una constante c por una variable X es igual al producto de esta constante por la desviación estándar de la variable, es decir:

$$s_{c \times X} = c \times s_X,$$

es decir, si a cada valor de la variable lo multiplicamos por una constante, entonces la nueva desviación estándar simplemente se obtiene multiplicando la desviación estándar anterior por esta constante.

d) También es importante señalar, que la desviación estándar es sensible a la presencia de datos atípicos, esto proviene del hecho que su cálculo se realiza en base al promedio, la cual ya hemos discutido es altamente sensible ante la existencia de dichos datos.

5.2.2. Didáctico

Variabilidad

El pensamiento estadístico requiere abordar la omnipresencia de la variabilidad en los datos, tanto dentro de un grupo como entre grupos o de una muestra a otra en una misma estadística. La comprensión de esta variabilidad constituye el núcleo del razonamiento estadístico, como plantean Wild y Pfannkuch (1999b), ya que permite interpretar los fenómenos en términos de cambio, incertidumbre y diferencia.

Diversos autores han enfatizado la importancia de que los estudiantes aprendan a reconocer la variabilidad como una característica esencial de los datos y no como un obstáculo o “ruido” (Konold y Pollatsek (2002); Ben-Zvi (2004)). Desde una perspectiva didáctica, analizar las fuentes y tipos de variabilidad

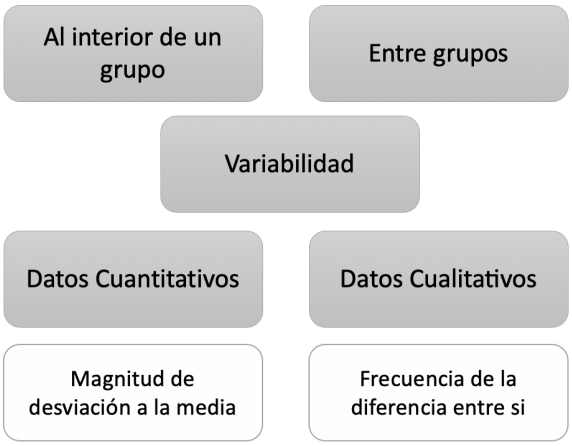


Figura 23: Variabilidad de los datos.

promueve el desarrollo de un pensamiento comparativo más sofisticado, que articula las nociones de centro, dispersión y forma de las distribuciones (Reading y Reid (2006)).

A partir de la observación de la variabilidad, y dada la necesidad de comparar distribuciones, emergen las medidas de dispersión tanto para datos cuantitativos como cualitativos. En los datos cuantitativos, la variabilidad se evidencia en las desviaciones de los valores respecto de la media, permitiendo introducir los conceptos de desviación estándar y varianza, tal como sostienen delMas et al. (2007). La Figura 24 ilustra distintas representaciones gráficas de la variabilidad en un histograma y en un gráfico temporal de barras, que facilitan la comprensión de cómo la dispersión puede representarse visualmente.

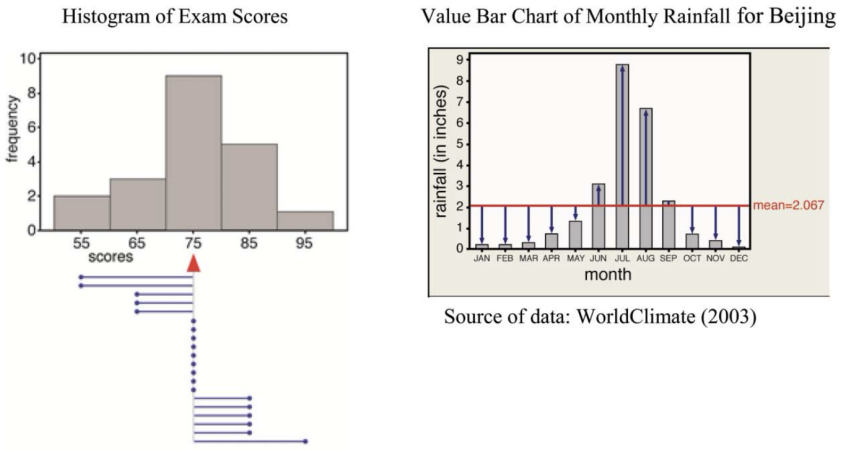


Figura 24: Comprensión de la variabilidad en gráficos cuantitativos.

En el primer ejemplo, la variabilidad puede estimarse mediante las distancias individuales de los datos respecto de la media, lo que conduce a las nociones de desviación y varianza. En el segundo, al analizar la distribución de la frecuencia de lluvias a lo largo de los 12 meses, se puede aplicar el modelo de igualdad de distancias para obtener una media de 2.067, evidenciando la dispersión temporal de los valores.

Por su parte, en los gráficos correspondientes a variables cualitativas, como el caso de los tipos de sangre en poblaciones japonesas y laponesas (Figura 25), la variabilidad se interpreta en términos del grado de diferencia o diversidad entre categorías (Kader y Perry (2007); Perry y Kader (2005)). En este ejemplo, el gráfico de los japoneses muestra una mayor variabilidad, ya que la distribución de frecuencias es más heterogénea. En los datos categóricos, por tanto, la variabilidad se entiende como el grado en que las frecuencias de las categorías difieren entre sí.

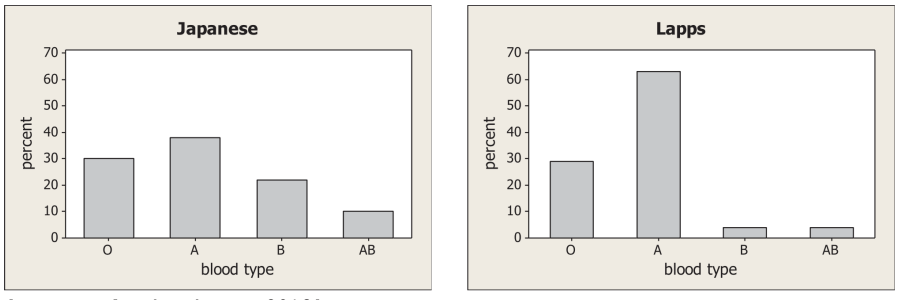


Figura 25: Comprensión de la variabilidad en gráficos cualitativos.

De acuerdo con Ben-Zvi y Arcavi (2001), la articulación de múltiples representaciones —gráficas, tabulares y simbólicas— permite a los estudiantes construir una visión global de los datos y desarrollar conexiones conceptuales entre las distintas formas de representar la variabilidad. En este sentido, el enfoque didáctico debe promover experiencias que permitan contrastar explícitamente la variabilidad cuantitativa, entendida en términos de *cuánto* los valores se desvían de la media, y la variabilidad categórica, comprendida en términos de *con qué frecuencia* los valores difieren entre sí (American Statistical Association y National Council of Teachers of Mathematics (2020)).

5.3. Orientaciones desde la implementación

A continuación se presenta una serie de recomendaciones para la implementación del problema, las cuales han surgido desde la experiencia de la aplicación de este en el aula.

1) Activación de conocimientos previos Antes de presentar el problema, se recomienda generar una instancia de conversación o actividades breves que permitan movilizar conocimientos sobre la media, la variabilidad y la dispersión. Algunas preguntas orientadoras son:
   a) ¿Qué información entrega la media sobre un conjunto de datos?
   b) ¿Es posible que dos conjuntos tengan la misma media, pero distribuciones distintas?
   c) ¿Qué entendemos por variabilidad o dispersión?
   d) ¿Qué gráficos nos ayudan a reconocer si los datos están más o menos dispersos?

   Esta etapa tiene como principal objetivo que los estudiantes reconozcan las relaciones entre las medidas de tendencia central y la forma de la distribución, y comprendan que la media no describe por sí sola la totalidad del comportamiento de los datos.
2) Promover la reflexión sobre el significado de la media. Se sugiere proponer situaciones simples, numéricas o visuales, que permitan discutir la media como punto de equilibrio. Por ejemplo, representar valores en una recta numérica o comparar dos pequeños conjuntos de datos con igual media. A partir de ello, se pueden guiar la reflexión a través de preguntas como:
   a) ¿Qué implica que la media esté “al centro” de la distribución?
   b) ¿Qué sucede con la media si algunos datos se alejan mucho del resto?
   c) ¿Cómo influye la dispersión en la representatividad de la media?

   Estas preguntas ayudan a que los estudiantes construyan una imagen mental de la media como una medida que depende del conjunto completo y no de valores aislados.
3) Fomentar la conexión entre representaciones. Antes de resolver el problema, es útil que los estudiantes relacionen representaciones gráficas y numéricas. Se pueden explorar distribuciones mostradas en gráficos de puntos o histogramas y estimar intuitivamente su media y su dispersión. Algunas preguntas orientadoras son:
   a) ¿Dónde ubicarías la media en este gráfico?
   b) ¿Qué tan “extendidos” o “concentrados” están los datos?

c) ¿Podría haber otra distribución con la misma media pero diferente forma? Esta pregunta hace alusión a que varios conjuntos de datos pueden tener la misma media, sin embargo, algunos pueden presentar mayor o menor dispersión que otros.

Este trabajo promueve la transición entre lo visual y lo numérico, un paso clave para comprender la relación entre la media y la variabilidad.

4) Propiciar la comparación y la argumentación. Una vez que los estudiantes han discutido distintas distribuciones o ejemplos, se recomienda guiarlos hacia la comparación entre conjuntos con igual media y diferente dispersión. Se puede pedir que expliquen, con sus propias palabras o mediante gráficos, ¿qué distribución consideran “más representativa” del promedio? y ¿por qué?
   Este tipo de intercambio fortalece la comprensión de la media como una medida sensible a los valores extremos y como un resumen que requiere interpretarse junto con la variabilidad.
5) Anticipar posibles dificultades conceptuales Es frecuente que los estudiantes:
   - Piensen que la media siempre corresponde a un valor existente en el conjunto.
   - No comprendan que la media se modifica con cada cambio en los datos.
   - Asocien la dispersión únicamente con la amplitud (rango) sin considerar cómo se distribuyen los valores.

   Por ello, se recomienda ofrecer experiencias de exploración con datos manipulables (por ejemplo, mover puntos en una recta) que permitan visualizar los efectos de los cambios sobre la media y la dispersión.

6. Problema 5

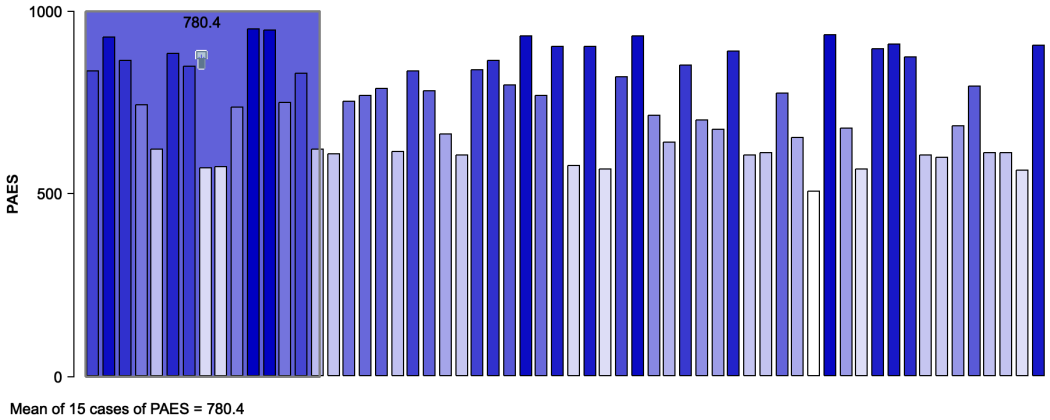
6.1. Planificación del Problema

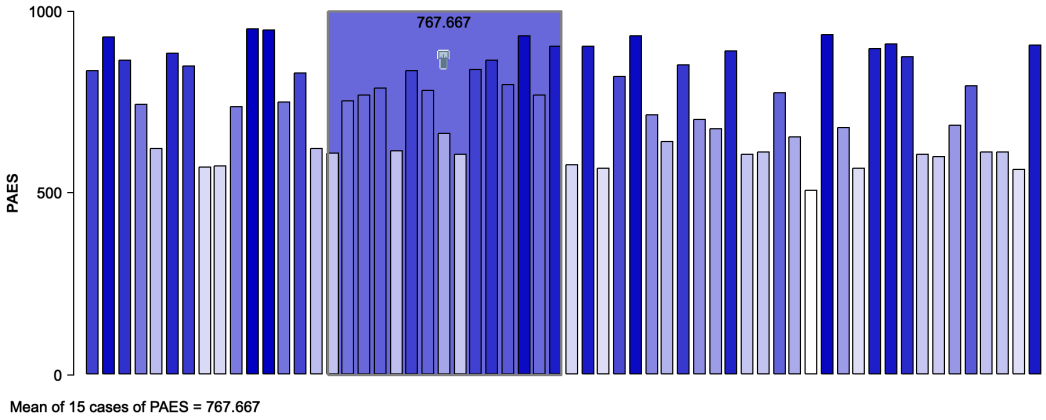
a) **Identificación curricular y de contenido del problema**

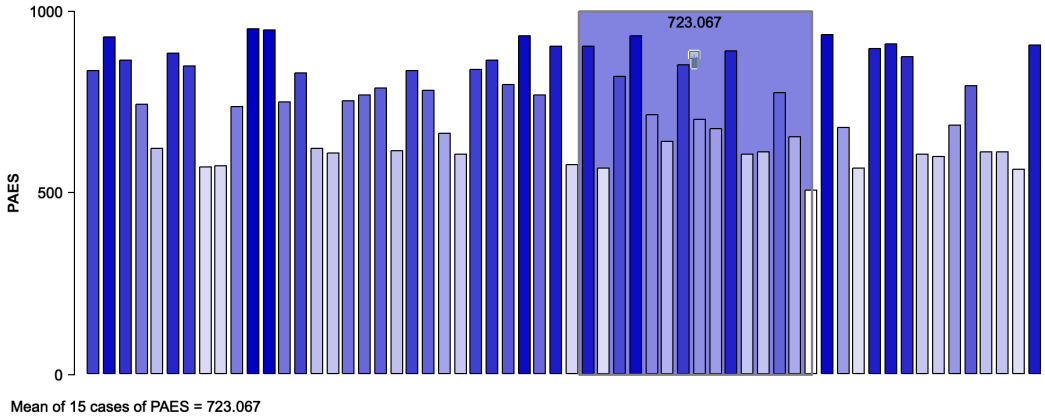
Objetivo de Aprendizaje del Currículo Escolar	Objetivo de Aprendizaje de la Actividad	Contenidos Estadísticos y de Probabilidad
OA 3. Argumentar inferencias acerca de parámetros (media y varianza) o características de una población, a partir de datos de una muestra aleatoria, bajo el supuesto de normalidad y aplicando procedimientos con base en intervalos de confianza o pruebas de hipótesis.	Establecer inferencias formales e informales a partir del análisis exploratorio de datos muestrales, indentificando las limitaciones y alcances para caracterizar los parámetros.	■ Media ■ Variabilidad ■ Dispersión ■ Inferencia Formal e Informal

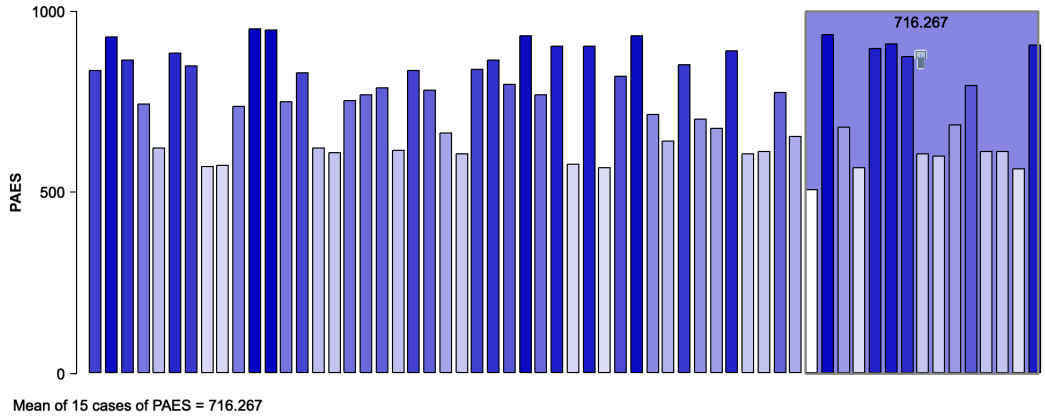
b) **Problema: La PAES**

Las siguientes imágenes visualizan la distribución de puntajes de dos cursos de estudiantes de cuarto medio en la prueba PAES. Cada cajón muestra la media muestral de 15 estudiantes.









¿Cuál es la media de los dos cursos de estudiantes?

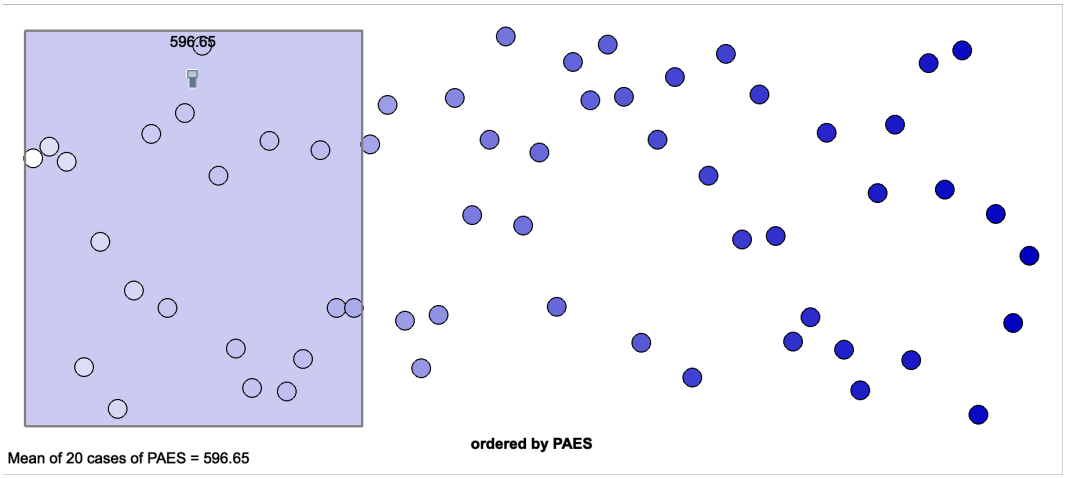
c) **Una solución del problema**

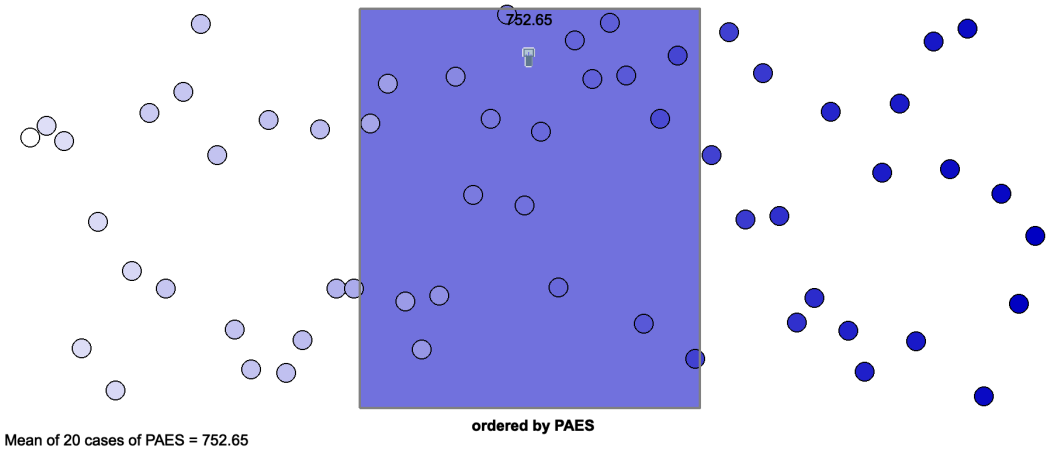
Para obtener el parámetro solicitado a partir de la información de los estadísticos muestrales, se puede considerar que los colores entregan información respecto de los puntajes obtenidos, por lo que se debería buscar una o un conjunto de muestras que fuera representativo de la distribución de colores de los dos cursos. Así, se puede seleccionar la segunda o tercera muestra (de izquierda a derecha) como cotas del parámetro. Esto porque la primera tiene colores muy marcados lo que se corresponde con la mayor frecuencia de puntajes altos; por otra parte la cuarta muestra considera colores claros, lo que se corresponde con la mayor frecuencia de puntajes bajos.

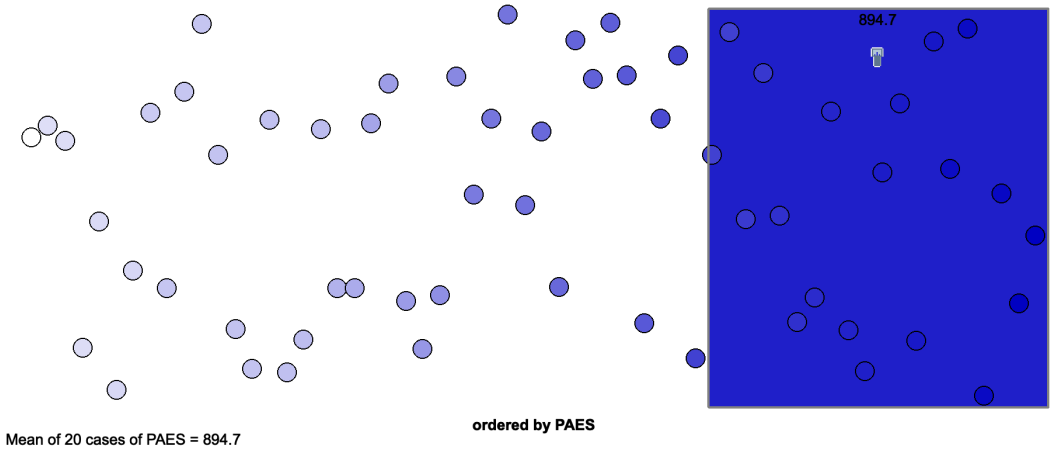
Otra forma de abordar el problema es observar que las muestras recorren toda la distribución de la población en estudio. Por lo que la combinación lineal de las medias, como estadísticos, permitiría obtener la media de la población, es decir, de los dos cursos.

d) **Una simplificación**

Las siguientes imágenes visualizan la distribución de puntajes de dos cursos de estudiantes de cuarto medio en la prueba PAES. Cada cajón muestra la media muestral de 20 estudiantes.







¿Cuál es la media de los dos cursos de estudiantes?

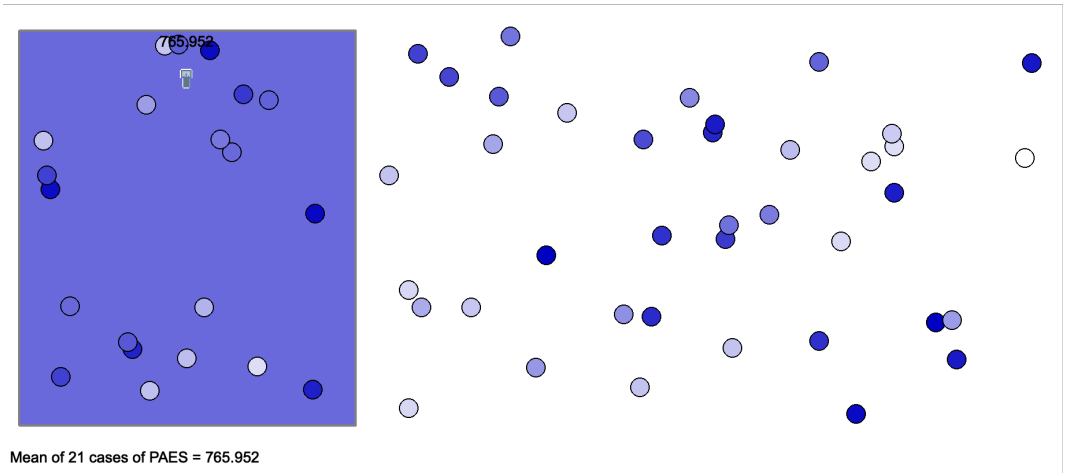
e) **Formas de consolidar el conocimiento o habilidades pretendidas (preguntas)**

Para consolidar el conocimiento realice las siguientes preguntas cuándo el grupo diga que tiene alguna respuesta a la situación:

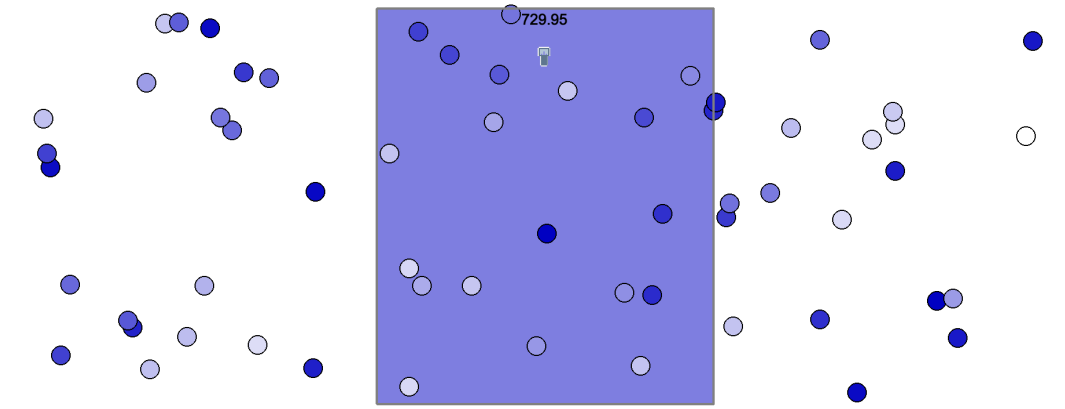
- ¿En qué intervalo podría estar la media de ambos cursos?
- ¿Cómo aseguran que en ese intervalo está la media de ambos cursos?
- ¿Cuál de los intervalos (estadístico) está más próximo a la distribución de la población?

f) **Una extensión**

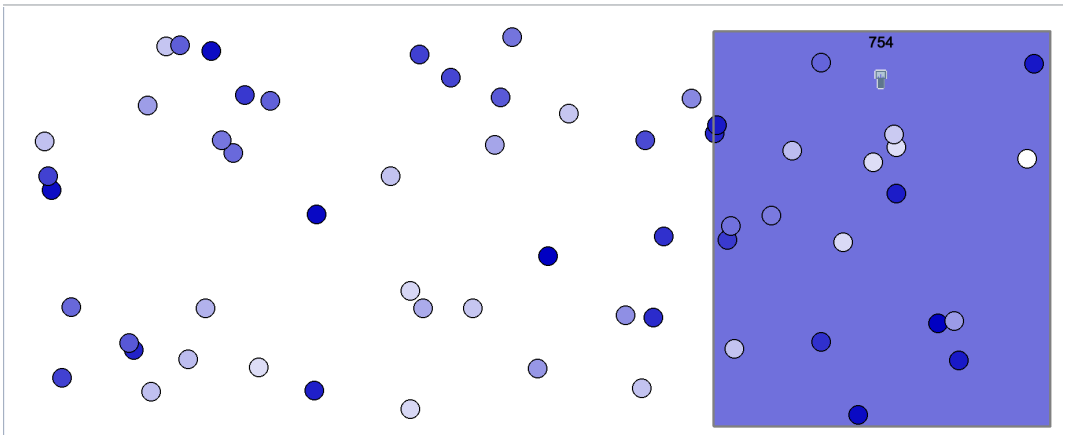
Las siguientes imágenes visualizan la distribución de puntajes de dos cursos de estudiantes de cuarto medio en la prueba PAES. Cada cajón muestra la media muestral de un grupo de estudiantes.



Mean of 21 cases of PAES = 765.952



Mean of 20 cases of PAES = 729.95



Mean of 20 cases of PAES = 754

¿Cuál es la media de los dos cursos de estudiantes?

g) **Plenaria** Para la plenaria se deben seleccionar respuestas que facilite la discusión entre los estudiantes en torno a: a) el conocimiento de las propiedades de la media al establecer estretegias de cálculo; b) la comprensión de la variabilidad al interior de las muestras; c) la estimación, como inferencia formal o informal de intervalos para justificar la presencia o no del parámetro poblacional. Las respuestas que pueden ser elegidas deben considerar las siguientes características:

- Mostrar un procedimiento algebraico para obtener la media de los cursos.
- Reconocer la distribución de la población para emparejarla con la de las muestras.

- Determinar intervalos para la media poblacional

6.2. Contenido para el profesor.

6.2.1. Disciplinar

Existen muchas definiciones para estadística, nosotros nos remitiremos a reconocerla como una ciencia que se encarga de sistematizar, recoger, ordenar y presentar los datos, relacionados con un determinado fenómeno de interés, el cual presenta variabilidad e incertidumbre. Deduce reglas y modela dichos fenómenos. La estadística permite hacer conjeturas sobre los fenómenos, permite concluir y tomar decisiones al respecto (generalizando sobre un conjunto mayor de datos). La Figura 26, esquematiza lo anterior, y muestra la relación directa del ciclo de investigación y por ende del método científico con la definición de estadística.

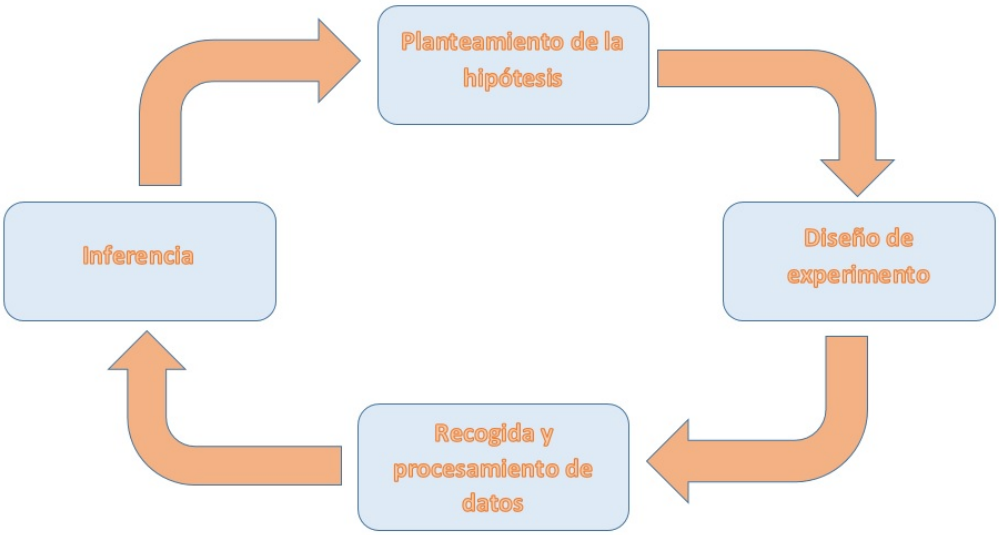


Figura 26: Ciclo de investigación

Analicemos cada etapa del ciclo de investigación y algunos de los conceptos claves en cada una de ellas:

- **Planteamiento de la(s) hipótesis:** Esta es la etapa de partida, en ella nos preguntamos ¿qué queremos investigar? o ¿qué queremos aprender de un determinado fenómeno?, ¿a qué o a quiénes involucra la investigación? Dichos planteamientos deben incluir una delimitación de la **población** de interés, la cual está formada por todos los individuos o sujetos (personas, objetos, animales, etc.) de las cuales se desea hacer un estudio y que poseen una característica en común, es decir, es el conjunto de todos los elementos de interés para un determinado problema (es hacia donde queremos inferir). Para formular estas hipótesis estadísticas se consideran algunas funciones definidas sobre valores numéricos de características medibles de la población objetivo, dichas funciones intentan resumir toda la información que hay en la población en unos pocos números y reciben el nombre de **parámetros**. Ejemplo: la altura media en cm. de un grupo de personas es un parámetro, al cual denotaremos como: μ. Otros parámetros de interés pueden ser la varianza, la desviación estándar (σ^2 y σ, respectivamente), etc.
- **Diseño del experimento:** En esta etapa se deben responder preguntas como: ¿qué características de la población nos interesan?, ¿cómo recogeremos la información requerida?, ¿qué individuos pertenecerán al estudio?, entre otras. La primera de estas preguntas hace alusión al concepto de **variable** ya definido en capítulos previos, tener claro las variables de interés es fundamental para planear las etapas de la investigación, está ligada directamente con el fenómeno que queremos analizar, estas deben estar definidas previo a la recogida de los datos (no po-

demos improvisar). En esta etapa también debemos responder a una pregunta transcendental: ¿se requiere una **muestra** o utilizaremos toda la población? Para responder a esta pregunta se deben tomar en cuenta aspectos tales como: tiempo, dinero, personal, distribución geográfica, etc, considerando que incluir a todos los individuos de una población de interés puede significar altos costos no sólo monetarios. Pero ¿qué implica tomar una muestra de la población?, lo primero es señalar que una **muestra** es un subconjunto de individuos de la población de interés, y lo segundo es que esta debe ser **representativa** de dicha población, esto implica que se debe velar para que la muestra conserve todas las características de la población. Para conseguir esto existen en estadística los llamados muestreos probabilísticos:

- **Muestreo aleatorio simple** (m.a.s), es el tipo de muestreo más popular, consiste en aleatorizar a todos los elementos de la población de manera tal que todos tengan la misma posibilidad de ser seleccionados en la muestra. Un ejemplo de m.a.s, es asignarle un número a cada individuo de la población (si es que se puede hacer una lista de todos ellos) y sortear dichos números hasta obtener el tamaño muestral deseado. El m.s.a tiene dos propiedades que lo convierten en el procedimiento utilizado por excelencia: a) todos los elementos de la población tienen la misma oportunidad de ser escogidos (es insesgado), b) la elección de un elemento de la población no influye sobre la elección de otro (independencia).
- **Muestreo sistemático**, en cual a grandes rasgos consiste en escoger un elemento al azar de la población y a partir de él de manera espaciada seleccionar a los restantes, por ejemplo, asignar números a los individuos, escoger al individuo en la posición 23 y a partir de él seleccionar a todos los invididuos cada 6 saltos, por lo tanto, los siguientes sujetos que formarán la muestra serán los que ocupan la posición: 29, 35, 41, etc. En este tipo de muestro se debe considerar que los individuos tienen un orden natural.
- **Muestreo estratificado**, para este tipo de muestreo se considera que la población de interés está subvidivida en grupos que no se intersectan, de acuerdo a una determinada variable de interés. Por ejemplo, se desea analizar si el tipo de dependencia del establecimiento de egreso de enseñanza media afecta en la proporción de estudiantes que ingresan a la universidad. Es claro que en este fenómeno los estratos a considerar son tres: Establecimientos Municipales o dependientes del Servicio Local de Educación (SLE), Establecimientos Subvencionados y Establecimientos Particulares. Cada uno de ellos forma un estrato, en el sentido que se asume que dentro de los estratos los individuos tienen características semejantes, estadísticamente se dice, son internamente homogéneos, mientras que entre los estratos se dice que son externamente heterogéneos. Una vez definidos los estratos se puede por ejemplo realizar un m.a.s dentro de cada uno de ellos para escoger la muestra, en cuyo caso se habla de muestreo estratificado aleatorio. Note que el parámetro de interés de este fenómeno son las proporciones.
- **Muestreo por conglomerados**, en este tipo de muestreo también se subdivide la población, pero en este caso los grupos a los que llamaremos conglomerados son heterogéneos internamente y homogéneos entre ellos. Por ejemplo, consideremos a todos los establecimientos educacionales que imparten enseñanza media en la cuidad de Concepción, suponga que estamos interesados en probar si el promedio de notas es igual en todos los niveles (en este caso el parámetro de interés es la media μ). Luego, los conglomerados serán los establecimientos, pues en cada uno de ellos se imparte de Primero a Cuarto Medio, lo que permite justificar la homogeneidad externa, la heterogeneidad dentro de cada conglomerado vendrá dado por el nivel. Luego, si sólo el interés del estudio se basa en establecer diferencias por nivel escolar, bastará con aleatorizar los conglomerados para obtener la muestra. Sin embargo, si se desean establecer diferencias en los promedios por nivel considerando por ejemplo el tipo de educación que imparte (Científico Humanista diurno, Científico Humanista vespertino, Técnico Profesional Comercial, Técnico Profesional Industrial, etc.) es posible que se requiera primero estratificar a la población. Es importante señalar entonces que la complejidad del muestreo depende exclusivamente del objetivo de investigación planteado el cual surge directamente de la formulación de la o las hipótesis.

Luego, en esta etapa del ciclo de investigación es relevante, decidir ¿cómo se tomarán los datos?, lo que involucra establecer el tipo de muestreo a utilizar. Otro punto importante de señalar en esta etapa, es que una vez que se decide obtener una muestra y no la población completa (CENSO), los cálculos que se realicen para probar las hipótesis no serán los parámetros como tal, sino que una estimación de ellos, los cuales se denominan **estadísticos**, y ellos estarán en base a la muestra. Precisamente en este punto radica la importancia de que la muestra sea representativa, debido a que si cumple con dicha condición, entonces los estadísticos serán una buena aproximación o estimación de los parámetros, lo que permitirá concluir en base a la hipótesis y tomar decisiones. Un ejemplo de estadístico es el promedio o media muestral denotado por $\overline{X}$, el cual es un estimador para la media poblacional (μ), en cuyo caso escribimos $\hat{\mu} = \overline{X}$. Por otro lado, si consideramos que σ^2 denota la varianza poblacional, un estimador para dicho parámetro es S^2, con lo cual se establece que $\hat{\sigma}^2 = S^2$, al cual denominamos varianza muestral. Cuando el muestreo no es probabilístico, por ejemplo, los datos se obtienen a partir de muestras con respuesta voluntaria, muestra por conveniencia, se dice que se produce un sesgo, lo cual provoca que se beneficien algunos resultados, en desmedro de otros, y por ende las conclusiones emitidas pueden resultar muy erróneas. Note que los estimadores son variables aleatorias, ya que corresponden a funciones definidas sobre los valores numéricos que describen una característica de la muestra. Por lo tanto, cambian cada vez que repetimos el muestreo y, en consecuencia, reflejan la variabilidad muestral. En este contexto, es importante distinguir:

  - En estadística descriptiva, simplemente calculamos los estadígrafos de la muestra, como $\bar{x},\ s^2,\ s$, los cuales son valores numéricos concretos.
  - En inferencia estadística, cuando hablamos de estimadores, nos referimos a las variables aleatorias que generan dichos estadígrafos. Por convención, los estimadores se denotan con letras mayúsculas (por ejemplo, $\bar{X},\ S^2,\ S$), para remarcar que representan una variable aleatoria, no un valor fijo.

- **Recogida y procesamiento de datos:** Esta etapa implica el proceso de medir las variables de interés en la muestra, guardarla apropiadamente para poder ser resumirla, presentarla y procesarla, esto último implica la aplicación de las herramientas estadísticas que permitan dar respuesta a la pregunta de investigación y probar las hipótesis planteadas.
- **Inferencia:** Esta es la etapa donde se interpretan los resultados obtenidos, donde a partir de la información recopilada de la muestra se concluye o toman decisiones con respecto a la población. Para llegar a esta etapa es crucial disponer de una muestra obtenida de manera probabilística, de lo contrario, aunque no inútil sólo será un análisis que no permitirá inferir hacia la población. Por ejemplo, si estamos interesados en probar que los estudiantes de enseñanza media diagnosticados con hiperactividad tienen peores promedio de notas en matemáticas en la región del BioBio, que quienes no han sido diagnósticados con hiperactividad; y por conveniencia solo considera a los estudiantes de primero medio pertenecientes a dos establecimientos en específico, es claro que no se puede inferir a la población completa. Sin embargo, esto puedo ser considerado en estadística un estudio piloto, información que puede servir para futuros estudios al respecto. Es importante señalar que en esta etapa, independiente de si se se cumplió o no la hipótesis del investigador, sirve para plantear nuevas hipótesis y/o reformular las anteriores, de esta manera se puede partir nuevamente con el ciclo de investigación.

Tal como se dijo previamente para probar la o las hipótesis de interés se recurre a la selección de una muestra representativa de la población, de manera tal que se pueda obtener una “buena” estimación de los parámetros. A continuación analizaremos esta relación entre la media poblacional y la media muestral.

Como hemos discutido previamente, la media aritmética solo puede calcularse para variables cuantitativas, sean estas discretas o continuas. Se caracteriza por ser altamente sensible a la presencia de valores atípicos y, además, siempre se encuentra acotada entre el valor mínimo y el valor máximo de la variable de interés. En otras palabras, la media pertenece al intervalo $[\text{mín}(X), \text{máx}(X)]$.

Analicemos la siguiente situación, se desea determinar el promedio de notas en matemática de

los estudiantes de enseñanza media de un establecimiento en particular, considerando un tamaño poblacional de 300 estudiantes, la Figura 27 muestra la distribución de dichas observaciones donde cada una de ellas en realidad corresponde al promedio de notas de cada estudiante.

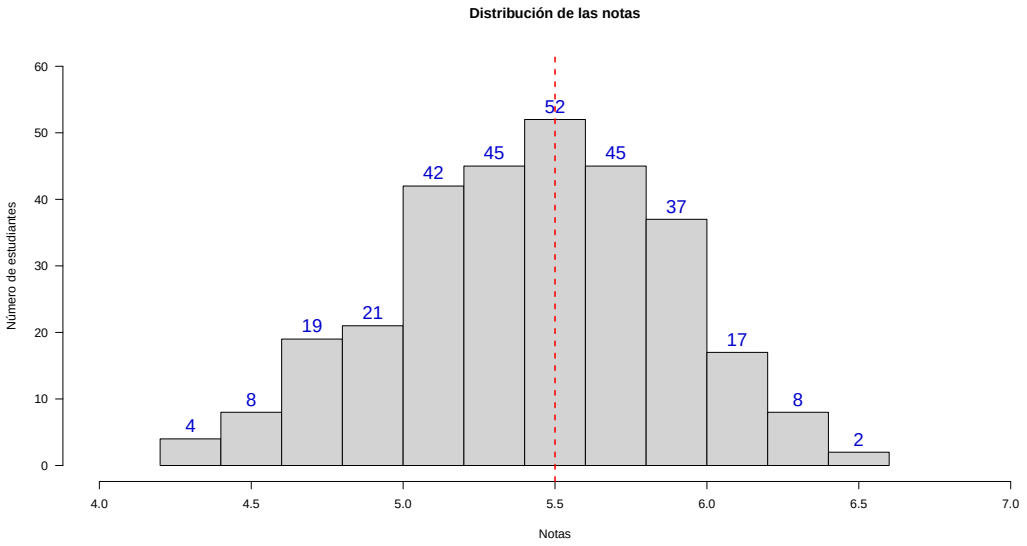


Figura 27: Distribución de los promedios de notas de 300 estudiantes

Supongamos que el establecimientos por medidas de confidencialidad de la información indica que no es posible disponer de la información de todos los estudiantes, sin embargo, entrega la información respectiva de 30 de estos alumnos los cuales son escogidos de manera aleatoria. La Figura 28 muestra la distribución de los promedios de notas de cuatro muestras de tamaño 30.

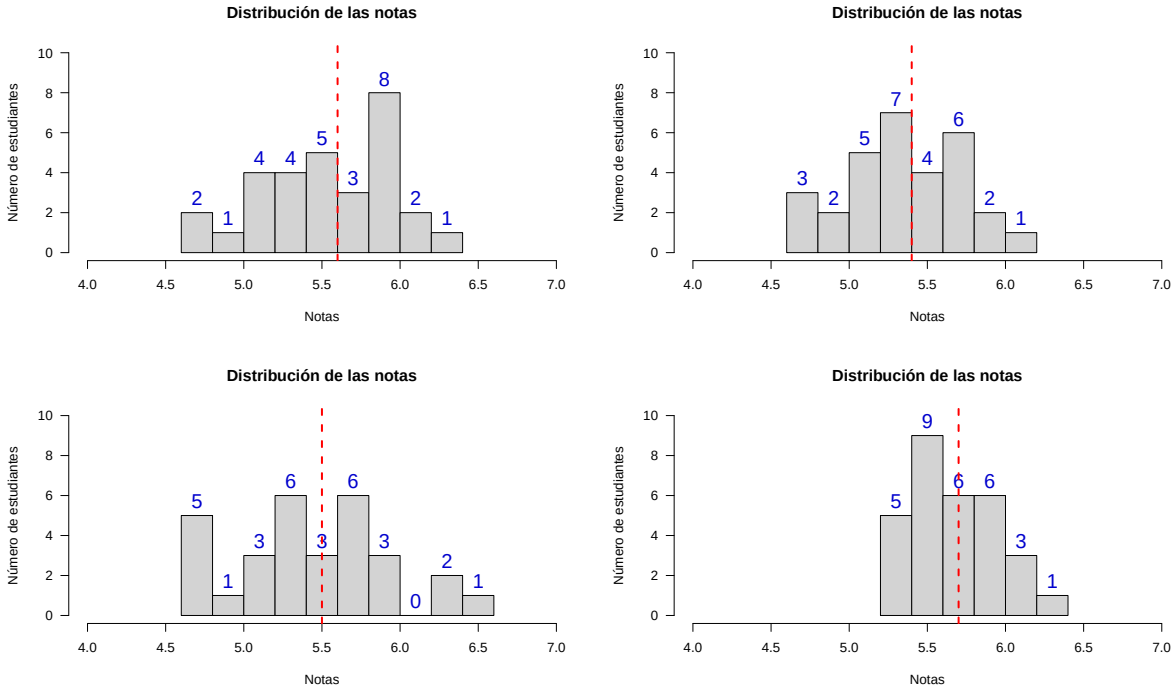


Figura 28: Distribución de los promedios de notas de dos grupos de 30 estudiantes

En cada figura, el promedio de los datos está representado por una línea vertical de color rojo. En la Figura 27, dicha línea corresponde al promedio de los promedios de nota de todos los estudiantes

$(N = 300)$. Como se calculó considerando a todos los elementos de la población, este valor es un parámetro, al que llamaremos media del promedio de notas o media poblacional. En cambio, en la Figura 28, cada línea roja representa el promedio obtenido a partir de una muestra aleatoria simple de $n = 30$ estudiantes. En este caso, se trata de un estadístico, al que denominaremos media muestral, ya que su valor depende de la muestra seleccionada. Así, la media muestral varía de una muestra a otra. En la figura se observan, por ejemplo, los siguientes valores de la media muestral: 5.6, 5.4, 5.5 y 5.7. Estos valores nos permiten estimar la media poblacional. Aunque se aproximan a ella, en general no coinciden exactamente. En este ejemplo, la media poblacional es $\mu = 5{,}5$. Esta diferencia entre el parámetro y el estadístico recibe el nombre de error de estimación.
Dado que el valor del estadístico varía de muestra en muestra decimos que existe variabilidad muestral, la cual es producto precisamente del proceso de muestreo. Si existe mucha variabilidad muestral, esto implica que el valor del estadístico cambia mucho entre muestras, lo cual no es bueno para el proceso de estimación. Sin embargo, si la muestra es obtenida de manera probabilística, la repetición de este procedimiento muchas veces considerando el mismo tamaño muestral mostrará un patrón.
La Figura 29 muestra la distribución del estadístico media muestral, considerando 100 muestras aleatorias de tamaño 30, luego 1000 muestras de tamaño 30, repitiendo para muestras de tamaño 100. Lógicamente este es sólo un ejercicios para ver como se comportan las medias muestrales, dado que el tamaño poblacional es 300, lo anterior no tendría sentido.

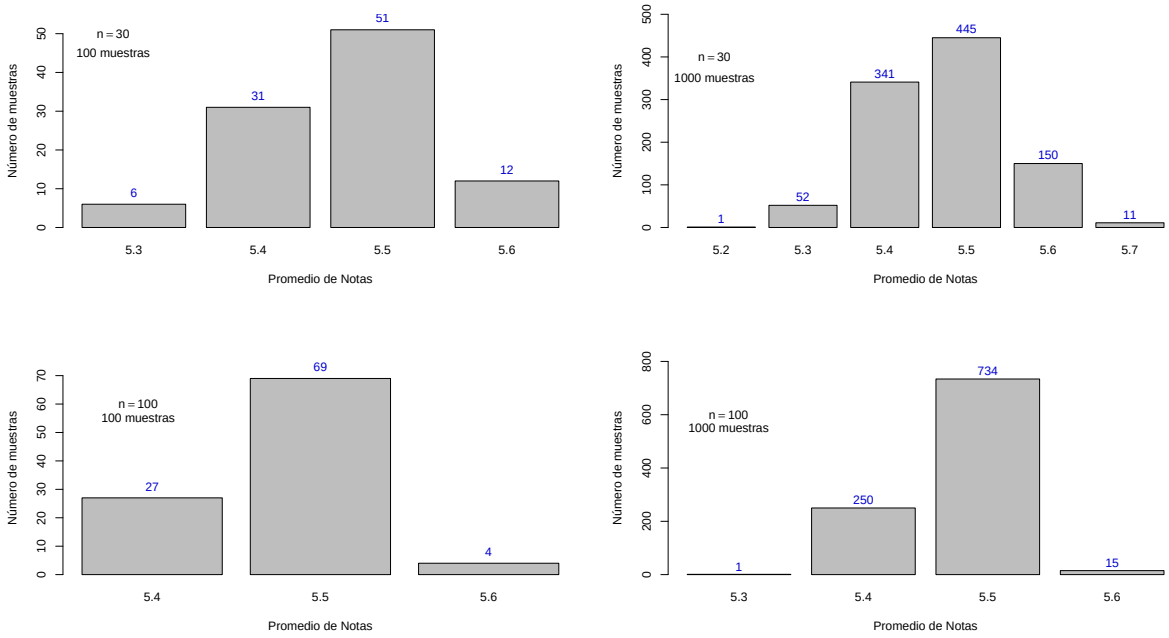


Figura 29: Distribución de los promedios muestrales para distintos tamaños muestrales.

A partir de estas figuras es posible concluir que, cuando el tamaño muestral es 30, si bien la frecuencia más alta corresponde al verdadero valor del parámetro en las dos figuras superiores; en la primera de ellas en el 51 $\%$ de las muestras efectivamente la estimación coincide con μ, sin embargo en la segunda figura pese a que se repitió más veces el proceso de muestreo este porcentaje decreció a 44.5 %. Repitiendo con un tamaño muestral igual a 100 y 1000 (las dos figuras inferiores), se observa que los porcentajes aumentan a 69 $\%$ y 73.4 $\%$, respectivamente. Es claro que las dos figuras superiores presentan mayor variabilidad muestral, sin embargo, debemos notar que en todos los casos es posible ver que la media muestral varía entre muestras (de lo contrario solo veríamos una única barra), y que la distribución de dichas medias está centrada en μ, esto implica que no hay sesgo y que es independiente del tamaño muestral.

6.2.2. Didáctico

Inferencia Estadística

La inferencia constituye el núcleo de la estadística, al ofrecer un medio para realizar afirmaciones fundamentadas bajo condiciones de incertidumbre cuando sólo se dispone de datos parciales. Harradine et al. (2011) la describen como el proceso mediante el cual se evalúa la fuerza de la evidencia sobre si un conjunto de observaciones es consistente o no con un mecanismo hipotético que podría haberlas generado. De modo complementario, Cobb y Moore (1997) la definen como el conjunto de métodos que permiten sacar conclusiones de los datos acerca de la población o el proceso del que provienen. Ambas perspectivas reconocen que la inferencia puede establecerse tanto de una muestra a una población como de una muestra a un proceso que la generó, incluso cuando la población no existe de forma tangible.

Rossman (2008) advierte que las definiciones comunes de la palabra *inferir* incluyen no sólo la conclusión obtenida, sino también las pruebas y el razonamiento que la sustentan. Sostiene que la inferencia implica ir más allá de los datos disponibles, ya sea para generalizar los resultados observados a un grupo mayor o para establecer relaciones más profundas entre variables, destacando el papel del azar como elemento *fundamental* para el razonamiento estadístico. En la misma línea, Makar y Rubin (2009) conceptualizan la inferencia estadística como una generalización probabilística —no determinista— que utiliza los datos como evidencia, integrando explícitamente la incertidumbre y la justificación basada en evidencia.

El argumento filosófico de Dewey sobre la función creativa de la inferencia y su vínculo con la valoración de la evidencia encuentra un paralelo en los enfoques exploratorios y confirmatorios de la inferencia estadística. La incorporación de formas exploratorias permite generar nuevos significados y sostener un proceso de indagación que equilibre la suspensión del juicio con la búsqueda activa de pruebas. En este sentido, Wild y Pfannkuch (1999b) destacan el papel del escepticismo productivo en la indagación estadística como motor del conocimiento. Desde una mirada contemporánea, Bakker y Derry (2011) proponen situar la inferencia en un marco filosófico que reconozca la interacción entre razonamiento, explicación y conocimiento contextual, argumentando que toda inferencia significativa debe integrar tanto el conocimiento estadístico como el conocimiento situado.

Asimismo, Cobb (2007) diferencian dos grandes tipos de inferencia: la que se establece de la muestra a la población y aquella que busca establecer relaciones causales a partir de un experimento. En la primera, se infiere sobre características poblacionales a partir de una muestra aleatoria; en la segunda, se infiere sobre causas o efectos mediante la asignación aleatoria de tratamientos o condiciones experimentales. En ambos casos, la inferencia estadística se apoya en la comprensión del azar, la variabilidad y la evidencia empírica.

Inferencia Estadística Informal

El análisis exploratorio de datos (AED), introducido por Tukey et al. (1977) y profundizado por Tukey (1980), propuso un cambio de paradigma al privilegiar la exploración y la visualización como medios para descubrir patrones e hipótesis significativas. Tukey concibió el AED no como un conjunto de técnicas, sino como una actitud flexible de indagación y descubrimiento, que antecede y complementa los procedimientos confirmatorios.

El auge de herramientas tecnológicas en las últimas décadas ha potenciado este enfoque, permitiendo desarrollar entornos de simulación y visualización que fortalecen la comprensión de la inferencia como proceso iterativo de construcción de evidencias, lo que constituye una base para la actual *ciencia de los datos*.

Zieffler et al. (2008) plantean la necesidad de articular el conocimiento informal de los estudiantes con la instrucción formal sobre inferencia estadística. A partir de este planteamiento, distintos marcos conceptuales han abordado el desarrollo de la inferencia estadística informal (IEI), entre ellos los de Ben-Zvi (2006), Makar y Rubin (2009), Pfannkuch (2006), Rubin et al. (2006), Rossman (2008) y Zieffler et al. (2008). Estos autores convergen en identificar cinco rasgos esenciales de la IEI: (a) la afirmación más allá de los datos, (b) la expresión de la incertidumbre, (c) el uso de los datos como evidencia, (d) la consideración del agregado y (e) la integración del contexto.

Afirmación más allá de los datos. Constituye el rasgo central de la inferencia estadística informal. Implica extender el razonamiento más allá de los datos disponibles para hacer afirmaciones sobre poblaciones o procesos no observados (Rossman (2008); Zieffler et al. (2008); Pfannkuch (2006)).
Expresión de la incertidumbre. La IEI incorpora el reconocimiento de la incertidumbre inherente a toda inferencia. Estas expresiones pueden ser cualitativas o informales, y evolucionan hacia un razonamiento probabilístico más preciso a medida que los estudiantes desarrollan experiencia (Ben-Zvi et al. (2012)).

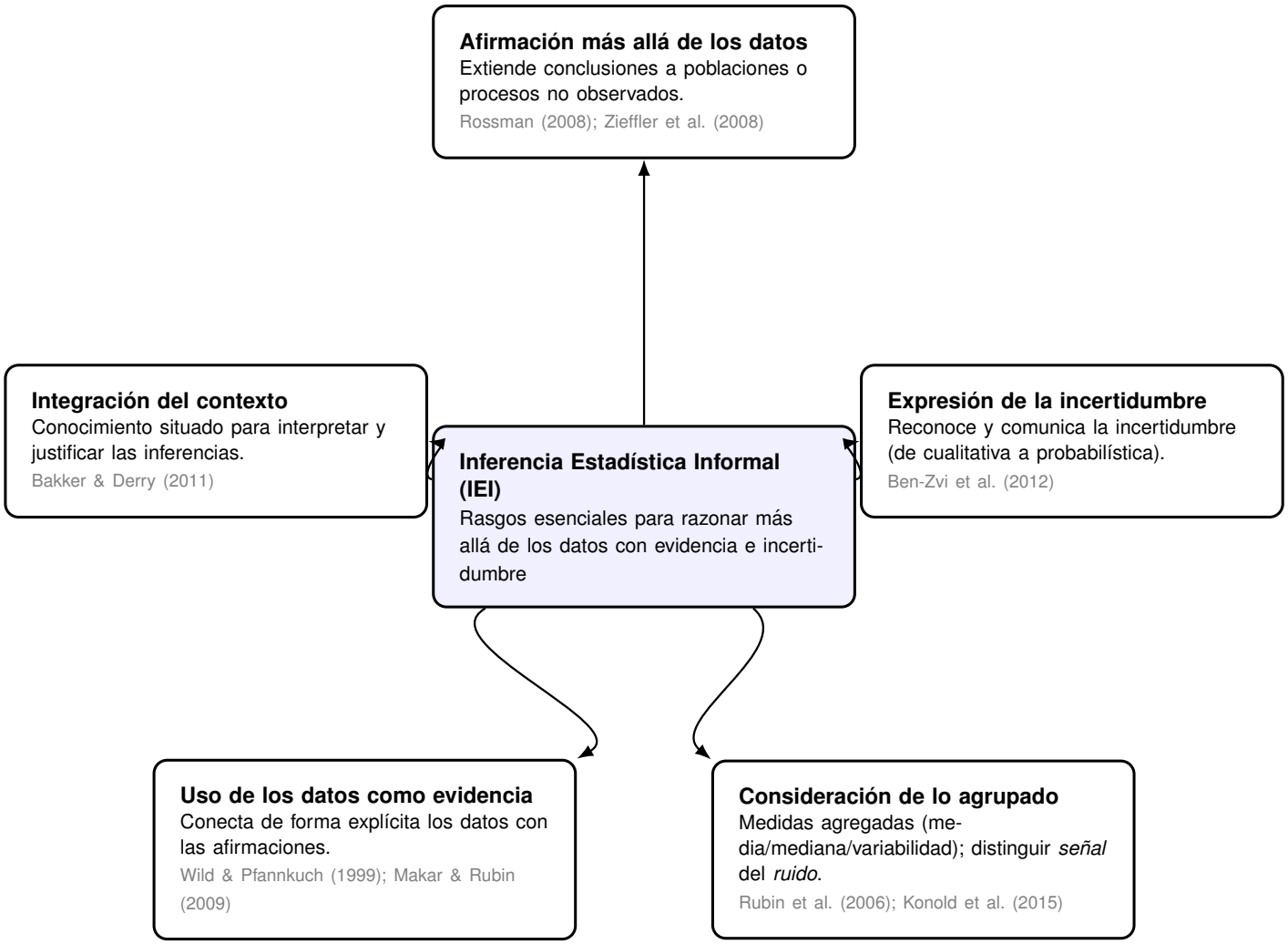


Figura 30: Cinco rasgos esenciales de la IEI: afirmación más allá de los datos, expresión de la incertidumbre, uso de los datos como evidencia, consideración de lo agrupado e integración del contexto.

Uso de los datos como evidencia. Para Makar y Rubin (2009), la inferencia requiere justificar las afirmaciones con base en evidencia empírica. Esta conexión entre datos y afirmaciones se construye gradualmente, y requiere explicitar cómo los datos sustentan las conclusiones (Wild y Pfannkuch (1999b); Fielding-Wells (2010); Hancock et al. (1992)).
Consideración de lo agrupado. La inferencia se apoya en medidas agregadas como la media o la mediana, que representan señales estables dentro de la variabilidad (Rubin et al. (2006); Aridor y Ben-Zvi (in press); Konold et al. (2015)). Esta distinción entre señal y ruido ayuda a los estudiantes a comprender cómo las propiedades colectivas de los datos reflejan regularidades subyacentes.
En conjunto, estas perspectivas configuran una visión didáctica de la inferencia estadística como un proceso de razonamiento que combina exploración, incertidumbre y evidencia, permitiendo transitar desde el pensamiento intuitivo hacia una comprensión formal y contextualizada de los procesos inferenciales.

6.3. Orientaciones desde la implementación

A continuación se presenta una serie de recomendaciones para la implementación del problema, las cuales han surgido desde la experiencia de la aplicación de este en el aula.
Lanzamiento: Presentar las imágenes al grupo y preguntar: "¿Qué información entrega el color de cada punto?" (tonos oscuros indican puntajes altos; claros, bajos). Luego: "Si cada cajón muestra la media de una muestra, ¿qué podríamos decir de la media de los dos cursos completos?"
Exploración guiada: Los estudiantes trabajan en parejas o tríos. Deben describir la mezcla de colores dentro de cada muestra, identificar la o las muestras más representativas de la distribución global y justificar su elección con base en la proporción visual de colores y cobertura espacial. Luego, proponen un intervalo razonable para la media poblacional usando dos estrategias: (a) emplear las medias muestrales seleccionadas como cotas inferior y superior; (b) combinar varias medias muestrales como promedio ponderado para estabilizar la estimación. Finalmente, escriben su estimación y tres razones que la respaldan.
Orientación docente: Explicar que las tres imágenes representan distintas muestras del mismo universo. Las diferencias en los promedios (aprox. 730, 754 y 766) muestran la variabilidad muestral. Discutir que la representatividad depende de cuán bien la muestra refleja la mezcla de colores del conjunto total.
Puesta en común: Cada grupo presenta su razonamiento. El docente puede ubicar las medias muestrales en una línea numérica y analizar su dispersión. Preguntas sugeridas: "¿Qué tan consistentes son las medias?" "¿Cómo influye la mezcla de colores en la estimación?" "¿Por qué promediar varias medias podría mejorar la precisión?"
Formalización: Reforzar las ideas centrales:

- La media poblacional puede estimarse a partir de medias muestrales.
- La variabilidad entre muestras es esperable: distintas muestras producen distintos promedios.
- Las muestras representativas reflejan la mezcla global de la población.
- Promediar varias medias muestrales reduce la influencia del azar.

Estrategias esperadas y dificultades: Se espera que los estudiantes seleccionen muestras visualmente representativas o combinen medias. Pueden confundirse si eligen muestras extremas (solo colores muy oscuros o muy claros), lo que introduce sesgo. También pueden ignorar el tamaño muestral; el docente debe enfatizar que muestras mayores tienden a generar medias más estables.
Preguntas para profundizar: "¿Qué evidencia visual te hace pensar que esta muestra es representativa?" "¿Cambiarías tu estimación si tuvieras más muestras?" "¿Cómo afectaría tu cálculo si promedias varias medias muestrales?" "¿Qué ocurriría si un curso tuviera muchos puntajes extremos?"

Como complemento de lo anterior se proponen los siguientes pasos:

1) Activación de conocimientos previos. Antes de presentar el problema, inicie preguntando a los estudiantes qué entienden por promedio o media, y si creen que siempre representa bien a un conjunto de datos. Algunas preguntas orientadoras son:
   a) ¿Qué representa la media en un conjunto de datos?
   b) ¿Dos grupos pueden tener la misma media, pero comportarse de forma diferente?
2) Promover la reflexión. A partir de ejemplos sencillos, motive a los estudiantes a reconocer que la media no describe por completo la distribución y que la variabilidad o dispersión también aportan información relevante sobre los datos. Para ello:
   a) Proponga comparar dos conjuntos de datos con igual media pero distinta dispersión, para que observen que la media por sí sola no describe bien la distribución.
   b) Introduzca la idea de variabilidad e invite a los estudiantes a pensar por qué, al tomar diferentes muestras de una misma población, las medias pueden variar.
   c) Proponga comparar distintas representaciones gráficas (por ejemplo, histogramas, boxplots) para que observen cómo cambia la forma de una distribución manteniendo igual media.

3) Propiciar la argumentación. Guíe la discusión hacia la idea de que, en contextos reales, solemos trabajar con muestras y que estas pueden ofrecer diferentes estimaciones de un mismo parámetro. Pregunte:
   a) ¿Por qué creen que las medias de distintas muestras extraídas de un mismo conjunto de datos no son exactamente iguales?
   b) ¿Qué tan distintas pueden ser y aún representar a la población?

   Estas preguntas permitirán introducir la noción de inferencia informal, vinculada con el razonamiento intuitivo sobre la representatividad de una muestra.
4) Anticipar posibles dificultades. Es esperable que algunos estudiantes:
   - Consideren la media como un valor absoluto que "describe a todos", sin reconocer la variabilidad entre muestras.
   - Interpreten la dispersión como un "error" en lugar de una característica natural de los datos.
   - No distingan entre una inferencia informal (basada en apreciaciones visuales o comparaciones directas) y una formal (basada en cálculos o modelos).

   Para abordar estas dificultades, se sugiere retomar representaciones visuales, promover la argumentación entre pares y enfatizar la conexión entre variabilidad muestral e incertidumbre inferencial.

7. Problema 6

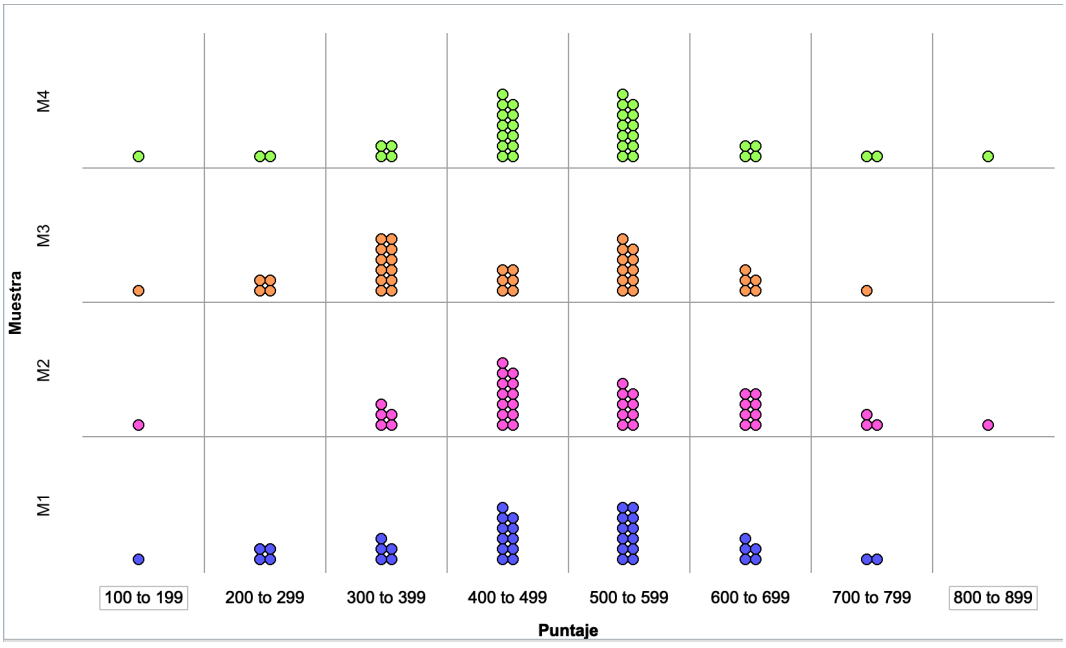
7.1. Planificación del Problema

a) **Identificación curricular y de contenido del problema**

Objetivo de Aprendizaje del Currículo Escolar	Objetivo de Aprendizaje de la Actividad	Contenidos Estadísticos y de Probabilidad
OA 3. Argumentar inferencias acerca de parámetros (media y varianza) o características de una población, a partir de datos de una muestra aleatoria, bajo el supuesto de normalidad y aplicando procedimientos con base en intervalos de confianza o pruebas de hipótesis.	Establecer inferencias formales e informales a partir del análisis exploratorio de datos muestrales, indentificando las limitaciones y alcances para caracterizar los parámetros de una población que se distribuye de forma Normal.	■ Media ■ Variabilidad ■ Dispersión ■ Inferencia Formal e Informal ■ Distribución Normal

b) **Problema: El parámetro**

La siguiente figura representa 4 muestras aleatorias que pertenecen a una población de puntajes PAES que tiene una desviación de 150 puntos. ¿Cuál puede ser la media de la población?



c) **Una solución del problema**

Cada una de las cuatro muestras (**M1**, **M2**, **M3** y **M4**) representa una selección aleatoria de puntajes provenientes de una población de resultados **PAES** con una desviación estándar conocida de 150 puntos. Al observar el gráfico, se aprecia que las muestras se concentran principalmente entre los **300 y 700 puntos**, con las medias de las muestras ubicadas aproximadamente en el centro de este rango.

- **M1** muestra una dispersión amplia, con puntajes entre 200 y 700, lo que sugiere una media cercana a 500 puntos.
- **M2** está algo más concentrada entre 400 y 600, con una media también próxima a 500–520 puntos.
- **M3** se distribuye mayormente entre 300 y 600, con una media aproximada de 480–500 puntos.
- **M4**, aunque más dispersa hacia los extremos, se centra también alrededor de 500 puntos.

Si las muestras son aleatorias e independientes, y se asume que provienen de una misma población con $\sigma = 150$, entonces la **mejor estimación puntual** de la media poblacional corresponde al promedio de las medias muestrales:

$$\hat{\mu} = \bar{X}_{población} \approx 500$$

En términos de inferencia estadística, las medias de las muestras tienden a agruparse alrededor de la verdadera media poblacional, lo que refleja el comportamiento esperado según el *teorema del límite central*. Por tanto, una estimación razonable para la media de la población es de aproximadamente **500 puntos**.

d) **Una simplificación**

Las siguientes 4 muestras pertenecen a una población de puntajes PAES cuya distribución es simétrica. ¿Cuál puede ser la media de la población de puntajes?

M1	391	269	396	518	295	588	543	364	534	612
	344	241	596	271	500	538	585	528	390	648
	659	468	486	487	716	669	625	457	556	560
M2	761	632	208	710	492	579	593	485	489	653
	731	484	577	532	472	482	652	470	194	471
	607	649	857	600	531	168	904	428	856	556
M3	581	592	592	246	555	645	691	466	452	723
	622	636	500	323	302	411	620	206	217	402
	250	434	569	257	542	782	499	458	571	458
M4	124	265	488	531	542	623	471	682	362	319
	559	363	633	550	474	623	558	433	535	597
	553	401	628	673	541	522	489	824	541	476

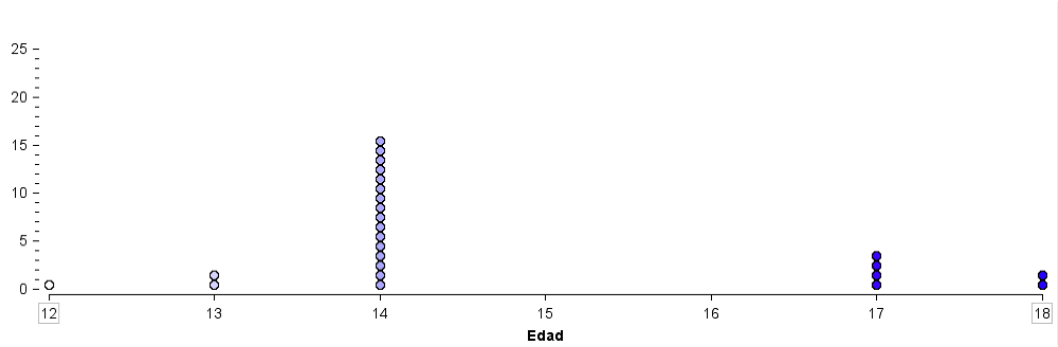
e) **Formas de consolidar el conocimiento o habilidades pretendidas (preguntas)**

Para consolidar el conocimiento realice las siguientes preguntas cuándo el grupo diga que tiene alguna respuesta a la situación:

- ¿Por qué están seguros que esa es la media?
- ¿Qué información reporta cada muestra para obtener la media poblacional?
- ¿Qué características proyectaron desde las muestras a la población de puntajes?
- ¿De qué forma utilizaron la información de la desviación?

f) **Una extensión**

La edad de los estudiantes de dos cursos se distribuye normal con media 15 años. Si entre los dos cursos suman 70 estudiantes, ¿cuál es la edad de los estudiantes que no están representados en el gráfico de puntos?



g) **Plenaria** Para la plenaria se deben seleccionar respuestas que faciliten la discusión entre los estudiantes en torno a: a) el conocimiento que aportan las muestras para caracterizar la población o; b) comprender el efecto de la variabilidad de la población de datos en las muestras, cuándo esta tiene una distribución normal o simétrica ; c) las características de una distribución normal. Las respuestas que pueden ser elegidas deben considerar las siguientes características:

- Mostrar un basado en el uso de variabilidad de la población en la determinación de la media poblacional.
- Reconocer las características de la población en base a las características de las muestras, con foco en la simetría

7.2. Contenido para el profesor.

7.2.1. Disciplinar

Con el objetivo de describir la distribución de un conjunto de datos, en secciones anteriores hemos revisado diversas herramientas como las tablas de frecuencias, histogramas, gráficos de caja (boxplots), polígonos de frecuencia y medidas de tendencia central, dispersión y posición. Aunque todas estas son útiles, presentan ciertas limitaciones. Por ejemplo, en las tablas de frecuencia para datos agrupados se pierde información, ya que dependen de las decisiones del analista —como la amplitud o el número de intervalos—, y en consecuencia el histograma asociado no siempre refleja con precisión la forma de la distribución. Del mismo modo, las medidas de tendencia central, dispersión y posición muestran únicamente aspectos parciales del comportamiento de los datos.

Por ello, es necesario contar con expresiones matemáticas que permitan describir el patrón general de los datos, siempre que estemos dispuestos a omitir observaciones inusuales o atípicas. A estas expresiones se les denomina **curvas de densidad**, y constituyen una herramienta fundamental para representar la distribución de un conjunto de datos, en especial cuando el tamaño muestral es grande. No obstante, debe tenerse presente que se trata de modelos matemáticos ideales: en la práctica, es muy difícil —cuando no imposible— que un conjunto de observaciones generado a partir de un proceso de muestreo se ajuste de manera perfecta a una curva de densidad.

Una de las curvas de densidad más conocidas es la curva de Gauss o curva Normal. Sin embargo, esta no puede aplicarse a cualquier conjunto de datos, ya que posee características específicas. Antes de profundizar en ellas, conviene destacar lo siguiente: la curva Normal se asocia a histogramas con forma de campana, donde los valores centrales de la variable concentran las frecuencias más altas, mientras que, al alejarnos de esta zona, las frecuencias disminuyen de manera simétrica. En otras palabras, las frecuencias muestran un comportamiento similar tanto a la izquierda como a la derecha respecto de los valores centrales.

Históricamente, esta curva fue introducida por Abraham de Moivre en 1733. Más tarde, recibió el nombre de campana de Gauss en honor a Carl Friedrich Gauss, quien en 1809 la utilizó para

describir los errores de observación cometidos por los astrónomos en mediciones repetidas. De forma paralela, Pierre-Simon Laplace la aplicó al análisis de errores en experimentos. A pesar de estos importantes aportes, no fue sino hasta 1889 cuando Francis Galton la denominó finalmente curva Normal.
Formalmente, diremos que una variable aleatoria X tiene distribución Normal con parámetros μ y σ^2, lo que denotamos como $X \sim N(\mu, \sigma^2)$, si su función de densidad está dada por:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \tag{3}$$

donde sus parámetros corresponden a la esperanza y varianza de X, respectivamente, es decir:

$$E(X) = \mu \qquad V(X) = \sigma^2$$

Las Figuras 31 y 32, muestran el rol de cada uno de estos parámetros, media y varianza, respectivamente. En la primera de ellas podemos observar que cuando la varianza es la misma ($\sigma^2 = 9$), las curvas presentan la misma forma acampanada, simétrica en torno a su respectiva media. También podemos ver que a medida que la media aumenta la curva respectiva se desplaza hacia la derecha, sin embargo, si pudiéramos sobreponerlas estas coincidirían, debido a esto se dice que la media es un parámetro de localización. Con respecto a la siguiente figura, es posible observar que manteniendo fija la media ($\mu = 17$), la curva a medida que aumenta la varianza se achata (sin perder su simetría en torno a la media), lo anterior se debe a que el área bajo la curva sigue siendo 1, por ende, si la varianza es mayor, implica mayor variabilidad entre los datos, es decir, existen datos más alejados de la media, los cuales la curva debe capturar (los datos no están tan concentrados). A la varianza por lo anterior, se le conoce como parámetro de escala.

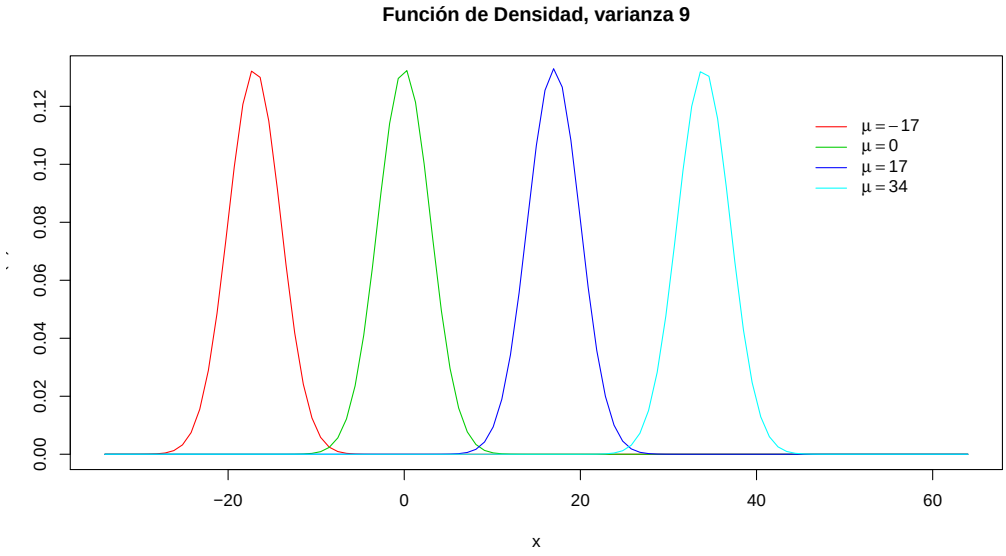


Figura 31: Función de densidad de la distribución Normal para distintos valores de μ.

La distribución normal tiene características importantes:

- El recorrido de la variable son todos los números reales.
- Su función de densidad de X es simétrica en torno a la media. Por lo cual, su media, moda y mediana coinciden. De ahí su forma de campana o cóncava hacia abajo.
- Sus colas decaen rápidamente, por lo que no se esperan valores muy alejados de la media.
- Posee dos puntos de inflexión, $(\mu \pm \sigma, f(\mu \pm \sigma))$, recordemos que un punto de inflexión es donde cambia la concavidad de una curva.

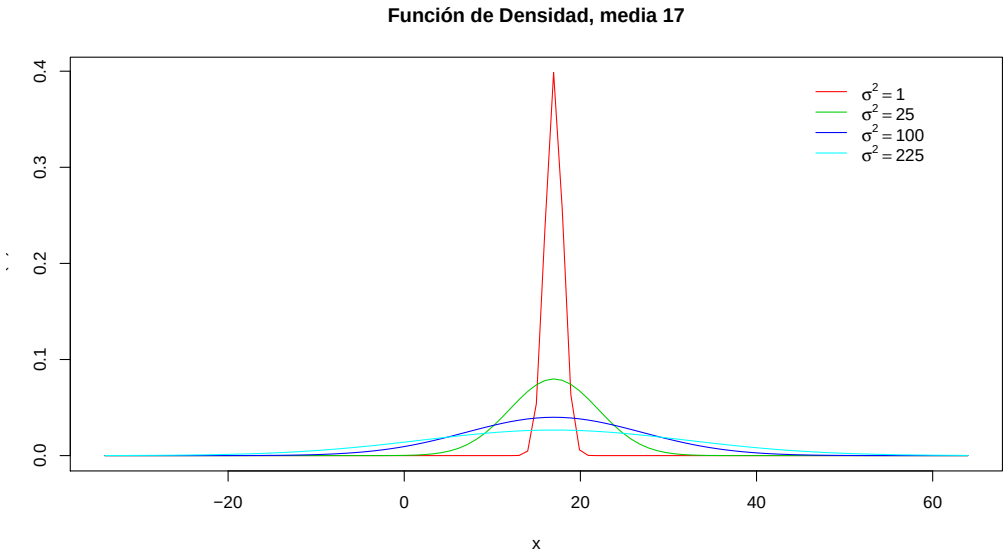


Figura 32: Función de densidad de la distribución Normal para distintos valores de σ^2.

La distribución Normal es muy importante en estadística, pues la mayoría de las técnicas estadísticas se basan que las muestras obtenidas provienen de esta distribución. Sin embargo, debemos tener en cuenta que rara vez los datos son perfectamente normales, no debemos idealizar a los histogramas, sino que debemos analizar si la curva Normal o densidad Normal nos entrega una buena aproximación al histograma de un determinado conjunto de datos y no esperar que calce perfecto. Las siguientes figuras (ver Figura 33) muestran diversos histogramas realizados a partir de datos simulados bajo la distribución Normal y distintos tamaños muestrales. En ellas se observa la media y la varianza de cada una y el tamaño muestral con la que fueron generadas, la curva roja representa la densidad bajo una distribución Normal con los parámetros señalados. En cada figura podemos notar que si bien la curva en cuestión no calza perfecta sobre los respectivos histogramas, esta describe casi perfecto el comportamiento de este, pues es capaz de captura prácticamente toda la distribución de los datos. Sin embargo, si disminuimos la amplitud de los intervalos la aproximación a la campana de Gauss será mejor.

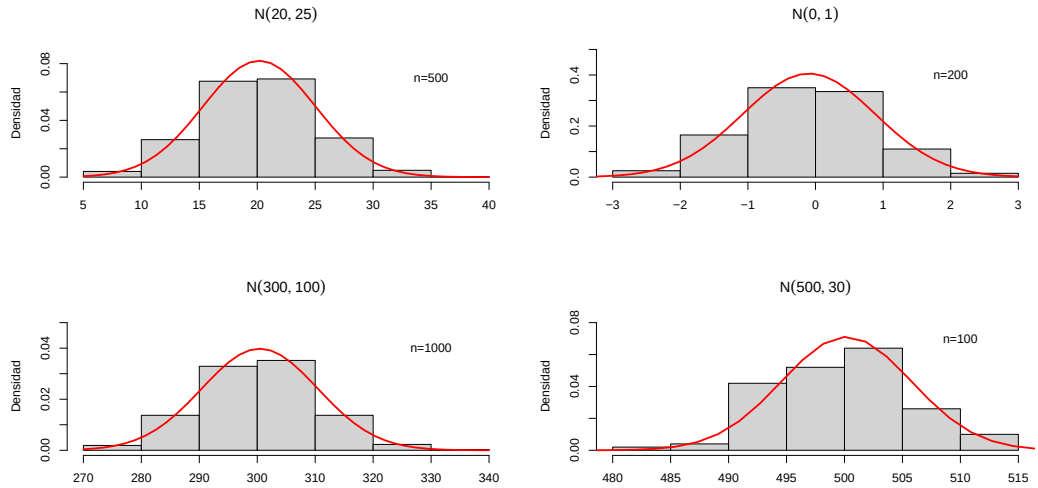


Figura 33: Función de densidad de distribuciones Normales.

La Figura 34 muestra un efecto contrario al caso anterior, los datos fueron generados con dis-

tribuciones distintas a la Normal, y se les ajustó una densidad Normal (línea roja). Es claro que intentar calzar a cada una de estas figuras la curva Normal no tiene sentido, pues los histogramas involucrados no son simétricos en torno a los valores centrales, ni aproximadamente, como en el caso anterior.

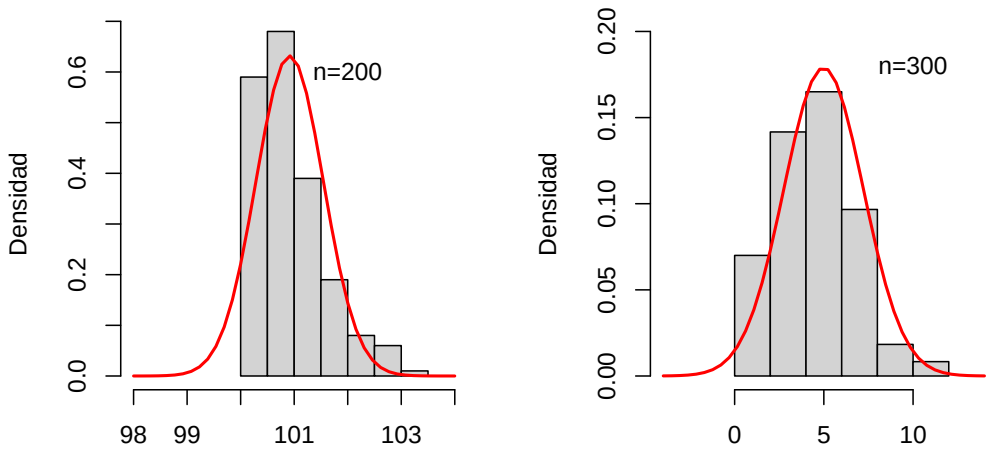


Figura 34: Función de densidad de distribuciones no Normal.

Es necesario notar que lo anterior fue concluido de manera gráfica y podríamos pensar que esto puede ser muy subjetivo, razón por la cual en estadística existen una serie de test que hacen más objetivo este análisis, sin embargo, estos escapan del objetivo de este trabajo.
Dadas las características de la distribución Normal se cumple que:

- El 68,26 $\%$ de los valores de una variable aleatoria Normal está dentro de más o menos una desviación estándar de su media.
- El 95,45 $\%$ de los valores de una variable aleatoria Normal está dentro de más o menos dos desviaciones estándar de su media.
- El 99,73 $\%$ de los valores de una variable aleatoria Normal está dentro de más o menos tres desviaciones estándar de su media.

Lo anterior gráficamente se puede ver como:
Luego, cuando se tiene evidencia de que los datos siguen una distribución Normal, basta con dos valores, los parámteros μ y σ^2, para determinar la distribución de todos los datos.
El cálculo de probabilidades es otra arista importante en Estadística, luego, contar con la expresión matemática que permita hacerlo es muy útil. Es sabido que en el caso de variables continuas las probabilidades se obtienen integrando con respecto a la variable de interés entre los valores requeridos, sin embargo, en el caso Normal integrar la expresión (3) no es una tarea fácil, razón por la cual se estandarizan los datos, de manera tal que expresión anterior no dependa de los parámetros. La manera de hacer esto es establecer la relación $Z = \dfrac{X - \mu}{\sigma}$.
Razón por la cual se habla de la distribución Normal estandarizada, cuya función de densidad está dada por:

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}},$$

y es denotada por: $Z \sim N(0, 1)$, es decir, posee media 0 y varianza 1, para calcular probabilidad bajo esta distribución solo basta usar la tabla de valores establecida.

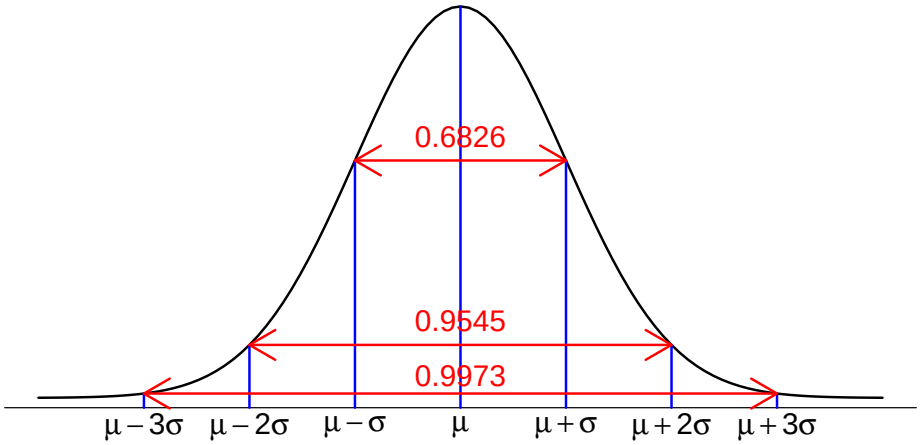


Figura 35: Función de densidad de la distribución Normal.

Ahora bien hemos hablado de la distribución Normal, asociándola al concepto de densidad, el cual en Estadística hace alusión a variables de tipo continuo, pero ¿sólo podemos usar la distribución normal en este tipo de variables?, la respuesta es NO. El Teorema del Límite Central, nos permite aproximar la distribución normal a una variable discreta, siempre que se cumplan determinadas condiciones, por ejemplo en el caso de la distribución binomial ($X \sim Binomial(n, p)$), si $n > 30$, $np > 5$ y $nq > 5$, es posible aproximar esta distribución por medio de la distribución Normal. Es necesario destacar que las condiciones de aproximación son requeridas para que se cumplan las características de la distribución normal, en algunos textos se sugieren por ejemplo tamaños muestrales mayores o que $np > 9$, etc.

7.2.2. Didáctico

La distribución normal ocupa un lugar central en la estadística y en la enseñanza de la probabilidad, no sólo por su presencia en fenómenos naturales y sociales, sino por su papel teórico como límite de las distribuciones muestrales de la media, tal como establece el *Teorema del Límite Central* (TLC). Este teorema sostiene que, "aunque los valores individuales provengan de una población muy poco normal, las medias muestrales tienden a tener una distribución normal" (Croucher, 2013, p. 467). Esta idea permite justificar el uso del modelo normal para realizar inferencias a partir de muestras, constituyéndose en un puente entre la variabilidad empírica y el razonamiento probabilístico formal.
En el caso chileno, el Ministerio de Educación de Chile (Mineduc) (2015) incorpora esta noción en el currículo escolar al proponer la esquematización del proceso de muestreo y de las distribuciones de probabilidad. Desde esta perspectiva, se espera que los estudiantes desarrollen una comprensión intuitiva y formal de las siguientes características de la distribución normal:

- La curva de la distribución de probabilidad normal tiene forma de campana.
- Es simétrica respecto a la media poblacional (μ).
- Las colas de la curva son asintóticas al eje horizontal.
- La distribución se describe mediante dos parámetros: la media (μ) y la desviación estándar (σ).

- El área bajo la curva entre dos puntos x_1 y x_2 representa la probabilidad de que una variable aleatoria X tome valores en ese intervalo.
- El área total bajo la curva es igual a uno. Por simetría, el área a cada lado de la media equivale a $0{,}5$ (50 %).

Una herramienta fundamental para favorecer la comprensión de este modelo es la **regla empírica** o *regla 68–95–99,7*, la cual describe la proporción de datos que se ubican dentro de 1, 2 y 3 desviaciones estándar de la media en una distribución normal estandarizada (z). Según McLeod (2019), esta regla constituye una vía accesible para que los estudiantes reconozcan la regularidad en la dispersión de los datos y vinculen el área bajo la curva con la noción de probabilidad.

- El 68 % de los datos cae dentro de una desviación estándar de la media.
- El 95 % se ubica dentro de dos desviaciones estándar.
- El 99,7 % se sitúa dentro de tres desviaciones estándar.

Factores que influyen en la comprensión de la distribución normal. La literatura ha documentado las dificultades persistentes en la comprensión del razonamiento estocástico y probabilístico. Wilensky (1995) y Wilensky (1997) introducen el concepto de *ansiedad epistemológica* para describir la sensación de confusión e indecisión que experimentan los estudiantes frente a problemas de incertidumbre, especialmente cuando reconocen la existencia de múltiples vías posibles de resolución pero carecen de recursos conceptuales para evaluarlas. Esta ansiedad se asocia con actitudes de evitación hacia la estadística y la probabilidad, reforzando la necesidad de enfoques didácticos que privilegien la exploración, la simulación y la interpretación visual de los fenómenos aleatorios (ver también Batanero et al. (2018)).

7.3. Orientaciones desde la implementación

La siguiente propuesta de orientaciones didácticas surge de la experiencia en el aula y busca guiar la implementación del problema "*Distribución de puntajes PAES*", centrado en la estimación de la media poblacional y la interpretación de la variabilidad muestral.

1) **Activar conocimientos previos.** Iniciar la actividad mediante preguntas sobre la dispersión observada en los puntajes y sobre qué representa "una media poblacional". Esto permite conectar con la idea de *tendencia central* y variabilidad.
2) **Explorar las muestras.** Presentar las cuatro muestras aleatorias de puntajes PAES (M1–M4) y solicitar a los estudiantes que estimen visualmente cuál podría ser la media de la población. Promover la argumentación basada en la posición de las muestras respecto al eje de los puntajes.
3) **Construir la inferencia.** Guiar la reflexión hacia el hecho de que las medias muestrales varían, pero tienden a agruparse en torno a un valor central, aproximando la media poblacional. Esta observación permite introducir la noción de *distribución muestral de la media*.
4) **Simular el proceso.** Utilizar herramientas digitales o planillas para simular múltiples muestras del mismo tamaño y graficar sus medias. Observar cómo la forma de la distribución se aproxima progresivamente a la normal, reforzando empíricamente el Teorema del Límite Central (Chance et al., 2002).
5) **Generalizar con la regla empírica.** Analizar cómo los valores de las medias se concentran dentro de uno o dos desvíos estándar de la media poblacional y conectar con la regla 68–95–99,7.
6) **Reflexionar sobre la incertidumbre.** Concluir destacando que toda inferencia a partir de una muestra conlleva incertidumbre, y que comprender la forma de la distribución normal permite cuantificar dicha incertidumbre de manera razonada.

Estas orientaciones integran la exploración visual, el razonamiento inductivo y la interpretación de modelos probabilísticos, buscando desarrollar una comprensión conceptual más profunda y menos algorítmica de la distribución normal.

8. Problema 7

8.1. Planificación del Problema

a) **Identificación curricular y de contenido del problema**

Objetivo de Aprendizaje del Currículo Escolar	Objetivo de Aprendizaje de la Actividad	Contenidos Estadísticos y de Probabilidad
OA 4. Argumentar inferencias acerca de parámetros (media y varianza) o características de una población, a partir de datos de una muestra aleatoria, bajo el supuesto de normalidad y aplicando procedimientos con base en intervalos de confianza o pruebas de hipótesis.	Reconocer las condiciones para establecer inferencias acerca de parámetros bajo el supuesto de normalidad, usando intervalos de confianza.	■ Media ■ Variabilidad ■ Inferencia Formal e Informal ■ Distribución Normal ■ Intervalo de confianza

b) **Problema: La línea de Maradona**
El Sernac está inspeccionando las bolsas de azúcar producidas por la firma ACME-Maradona, que supuestamente traen un kilo cada una. Se toma una muestra aleatoria simple de 850 bolsas de una partida. El peso promedio resulta ser 996 g. ¿Cuál es el intervalo de confianza para el peso promedio real de las bolsas?

c) **Una solución del problema**
Datos: Muestra aleatoria simple de $n = 850$ bolsas; media muestral $\bar{x} = 996$ g. Se desea un IC para la media poblacional μ del peso de las bolsas.

(A) Desviación estándar *desconocida* (caso habitual). Con n grande, el IC del $95\,\%$ para μ es

$$\mu \in \bar{x} \pm z_{0{,}975}\,\frac{s}{\sqrt{n}} \qquad \Rightarrow \qquad \mu \in 996 \pm 1{,}96\,\frac{s}{\sqrt{850}},$$

donde s es la desviación estándar *muestral* (estimada con los 850 pesos).

Interpretación: una vez calculado s, se reemplaza en la expresión y se obtiene el IC.

Ejemplo numérico (ilustrativo): si $s = 24$ g, entonces

$$\text{ME} = 1{,}96 \cdot \frac{24}{\sqrt{850}} \approx 1{,}61 \text{ g} \quad \Rightarrow \quad \mu \in (996 - 1{,}61,\ 996 + 1{,}61) = (994{,}39,\ 997{,}61) \text{ g}.$$

Si $s = 30$ g,

$$\text{ME} = 1{,}96 \cdot \frac{30}{\sqrt{850}} \approx 2{,}02 \text{ g} \quad \Rightarrow \quad \mu \in (993{,}98,\ 998{,}02) \text{ g}.$$

(B) Desviación estándar *conocida* (histórica del proceso). Si existe una desviación poblacional conocida σ (p. ej., de control de proceso), use

$$\mu \in \bar{x} \pm z_{0{,}975}\,\frac{\sigma}{\sqrt{n}} \qquad \Rightarrow \qquad \mu \in 996 \pm 1{,}96\,\frac{\sigma}{\sqrt{850}}.$$

Hecho útil (¿entra 1000 g en el IC del 95 %?): Para que 1000 g quede *dentro* del IC del 95 %, se requiere que el margen de error sea al menos 4 g:

$$1{,}96\,\frac{s}{\sqrt{850}} \;\geq\; 4 \quad \Longleftrightarrow \quad s \;\geq\; \frac{4\sqrt{850}}{1{,}96} \;\approx\; 59{,}5 \text{ g}.$$

Es decir, salvo que la variabilidad sea muy grande ($s \gtrsim 60$ g), el valor objetivo de 1000 g no quedará cubierto por el IC del 95 % centrado en 996 g.

Conclusión: Con la media muestral de 996 g y $n = 850$, el IC del $95\,\%$ para la media poblacional es

$$\boxed{\mu \in 996 \pm 1{,}96\,\frac{s}{\sqrt{850}}}$$

(ó con σ si es conocida). Con valores de dispersión típicos de procesos de envasado (decenas de gramos), el intervalo queda claramente por *debajo* de 1000 g, lo que sugiere *subllenado* sistemático.

d) **Una simplificación**
El Sernac está inspeccionando las bolsas de azúcar producidas por la firma ACME-Maradoma, que supuestamente traen un kilo cada una. Se toma una muestra aleatoria simple de 850 bolsas de una partida. El peso promedio resulta ser 996 g, con una desviación típica de 50 g. ¿En que casos intervalo de confianza para el peso promedio real de las bolsas contiene el valor anunciado de 1000g?

e) **Formas de consolidar el conocimiento o habilidades pretendidas (preguntas)**
Para consolidar el conocimiento realice las siguientes preguntas cuándo el grupo diga que tiene alguna respuesta a la situación:

- ¿Por qué están seguros que esa es la media?
- ¿Qué información reporta cada muestra para obtener la media poblacional?
- ¿Qué características proyectaron desde las muestras a la población de puntajes?
- ¿De qué forma utilizaron la información de la desviación?

f) **Una extensión**
El Sernac está inspeccionando las bolsas de azúcar producidas por la firma ACME-Maradoma, que supuestamente traen un kilo cada una. Se toma una muestra aleatoria simple de bolsas de una partida. El peso promedio resulta ser 996 g. ¿Cuál es el intervalo de confianza para el peso promedio real de las bolsas al 90 %, al 95 % y al 99 %?

g) **Plenaria** Para la plenaria se deben seleccionar respuestas que faciliten la discusión entre los estudiantes en torno a: a) el conocimiento que aportan las muestras para caracterizar la población o; b) comprender el efecto de la variabilidad de la población de datos en las muestras, cuándo esta tiene una distribución normal o simétrica ; c) las caracteristicas de una distribución Normal. Las respuestas que pueden ser elegidas deben considerar las siguientes características:

- Mostrar un basado en el uso de variabilidad de la población en la determinación de la media poblacional.
- Reconocer las características de la población en base a las características de las muestras, con foco la simetría

8.2. Contenido para el profesor.

8.2.1. Disciplinar

Hasta ahora hemos visto que para obtener una estimación del valor de un parámetro basta con sustituir las observaciones de la muestra en el estimador encontrado. Sin embargo, este valor no necesariamente es el verdadero valor de dicho parámetro. ¿Por qué?
Pues debido a que la estimación puntual es producto de una m.a, imaginemos la siguiente situación. Supongamos que se toman 3 m.a de tamaño 2, a partir de los 5 primeros números naturales pares, sin reemplazo.

Por ejemplo:
m.a 1 formada por los números 2 y 4.
m.a 2 formada por los números 4 y 8.
m.a 3 formada por los números 6 y 10.

Es evidente que el promedio de la muestra no necesariamente coincide con la media poblacional, que en este caso sabemos es $\mu = 6$. Por lo tanto, podemos afirmar que toda estimación está sujeta a un error de estimación.
Por esta razón es mucho más adecuado construir intervalos de confianza (IC) en donde el parámetro que se estima está contenido con cierta probabilidad $(1-\alpha)$, a la cual se le denomina **nivel de confianza** en él. Por lo que en lugar de estimar un parámetro mediante un único valor, estimamos un rango de valores para dicho parámetro.
Luego, diremos que dado $X_1, X_2, ..., X_n$ una m.a de una población X, cuya función de probabilidad recordemos está dada por $f_X(x;\theta) = f_X(x)$, y de donde se tendrá que:

$$P(\hat{\theta}_1 \leq \theta \leq \hat{\theta}_2) = 1 - \alpha.$$

Luego, un intervalo de confianza bilateral para θ estará dado por:

$$[\hat{\theta}_1, \hat{\theta}_2],$$

donde $\hat{\theta}_1$ y $\hat{\theta}_2$ son dos estadísticos denominados límite inferior y límite superior de confianza para θ, respectivamente.
La elección del nivel de confianza depende del investigador, pero los valores más utilizados son:

$$0,90 - 0,95 - 0,975 - 0,98 - 0,99.$$

Un método útil para determinar intervalos de confianza es el método del pivote, el cual debe tener las siguientes características.

a) Debe ser función de las mediciones de la muestra y del parámetro desconocido θ.
b) Su función de probabilidad y/o densidad, no debe depender del parámetro θ.

A continuación obtendremos el intervalo de confianza para la media de una distribución Normal. Sea $X_1, X_2, ..., X_n$ una m.a de $X \sim N(\mu, \sigma^2)$, nuestro interés es determinar un IC para μ $(\theta = \mu)$. Suponiendo que σ^2 es conocido, debemos considerar,

- Primero escoger el nivel de confianza $(1 - \alpha)$.
- Como lo que deseamos es un intervalo para μ y sabemos que la media muestral, $\bar{X}$, es adecuada para estimar μ, por lo tanto podemos usar la distribución muestral de $\bar{X}$ para establecer un IC para μ.

Entonces, lo primero que debemos hacer es obtener la distribución de $\bar{X}$, como $X_1, X_2, ..., X_n$ es una m.a de $X \sim N(\mu, \sigma^2)$, implica que $X_1, X_2, ..., X_n$ son independientes y que cada $X_i \sim N(\mu, \sigma^2)$, lo que implica que $E(X_i) = \mu$ y $Var(X_i) = \sigma^2$. Luego, por propiedad reproductiva de la distribución Normal se tiene que la suma de variables con distribución Normal distribuye

Normal, y que su media muestral también (para una demostración formal se puede usar la función generadora de momentos, temática que escapa del objetivo de este material). Sólo nos falta determinar la esperanza y la varianza de esta variable:

$$
\begin{aligned}
E(\bar{X}) &= E\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) && \text{por definición de } \bar{X} \\
&= \frac{1}{n} E(X_1 + X_2 + \cdots + X_n) && \text{prop. de esperanza } (n \text{ constante}) \\
&= \frac{1}{n}\big[E(X_1) + E(X_2) + \cdots + E(X_n)\big] && \text{por linealidad de la esperanza} \\
&= \frac{1}{n}[\mu + \mu + \cdots + \mu] && \text{porque } E(X_i) = \mu,\ i = 1, ..., n \\
&= \frac{n\mu}{n} \\
&= \mu.
\end{aligned}
$$

$$
\begin{aligned}
Var(\bar{X}) &= Var\left(\frac{\sum_{i=1}^{n} X_i}{n}\right), && \text{por definición de } \bar{X} \\
&= \frac{1}{n^2} Var(X_1 + X_2 + ...X_n), && \text{prop. de varianza } (n \text{ constante}) \\
&= \frac{1}{n^2}[Var(X_1) + Var(X_2) + ... + Var(X_n)], && \text{porque las } X_i \text{ son independientes} \\
&= \frac{1}{n^2}[\sigma^2 + \sigma^2 + ... + \sigma^2], && \text{porque } Var(X_i) = \sigma^2,\ i = 1, ..., n \\
&= \frac{n\sigma^2}{n^2} \\
&= \sigma^2/n. && \text{simplificando}
\end{aligned}
$$

Luego, $\bar{X} \sim N(\mu, \sigma^2/n)$, de donde al estandarizar obtenemos:

$$Z = \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \sim N(0, 1).$$

Note que la cantidad anterior cumple con las condiciones para ser pivote, pues: 1) Z es función de la muestra y del parámetro μ, y 2) su función de densidad no depende del parámetro a estimar. Luego, usando el método del pivote tenemos:

$$P(z_1 \leq Z \leq z_2) = 1 - \alpha,$$

donde z_1 y z_2 corresponden a dos percentiles particulares de la distribución Normal, tal que entre ellos se encuentra μ con un nivel de confianza de $1 - \alpha$.
En la Figura 36, se puede observar de manera gráfica lo anterior, en ella se representa que z_1 corresponde al percentil que acumula un área igual a $\alpha/2$ y que z_2 es el percentil que acumula un área igual a $1 - \alpha/2$.

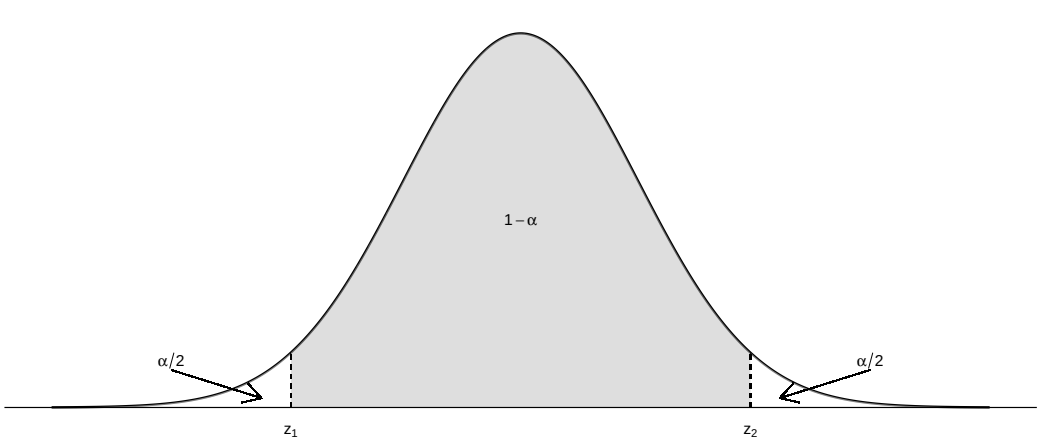


Figura 36: Curva Normal estándar con intervalo de confianza $1-\alpha$ y colas $1-\alpha/2$.

Reemplazando y gracias a la simetría de la distribución Normal, donde $z_{\alpha/2} = -z_{1-\alpha/2}$ tenemos:

$$P\left(-z_{1-\alpha/2} \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq z_{1-\alpha/2}\right) = 1-\alpha$$

Despejando adecuadamente obtenemos:

$$P\left(\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

$$\therefore \quad \mu \quad \in \quad \left[\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

Es un intervalo de confianza con un nivel de confianza del $(1-\alpha)100\,\%$ para μ. Note que los extremos del intervalo son aleatorios, pues dependen de la m.a, y como tal están centrados en $\bar{X}$ no en el parámetro μ.

Interpretación gráfica de los intervalos

La Figura 37, muestra 100 intervalos de confianza del 95 $\%$ para μ, en ella se observa que 5 de estos intervalos (los de color rojo) no contienen al verdadero valor del parámetro que estamos estimando. Debemos considerar entonces que cuando calculamos intervalos de confianza a partir de una muestra, lo que obtenemos es uno de los tantos que podríamos construir, dado que ellos depende de los valores recopilados. Luego, nosotros confiamos que el intervalo construido contenga al verdadero valor nuestro parámetro de interés, pero no estamos 100 $\%$ seguros que esto sea así. En el caso de la figura señalada, el 95 $\%$ de las veces será así.

Análogamente, se verían las figuras correspondientes al 90 $\%$ de nivel de confianza, donde serían 10 los intervalos de color rojo, para $99\,\%$, donde sólo sería un IC de color rojo, etc.

Como podemos ver la elección del nivel de confianza es un factor que afecta la longitud del intervalo construido, analicémoslo, manteniendo fijos el tamaño muestral y la desviación estándar:

Note que a medida que el nivel de confianza aumenta, la longitud del intervalo también lo hace, esto es evidente, pues cuando aumentamos $(1-\alpha)$ estamos intentando que con mayor confianza el intervalo construido contenga el verdadero valor del parámetro, en otras palabras estamos dispuestos a equivocarnos menos (ver Tabla 13). Observe que esto se debe a que el valor del percentil aumenta.

De la misma manera, si mantenemos fijo el nivel de confianza y la desviación estándar, y aumentamos el tamaño de la muestra, podemos concluir que los intervalos tendrán menor longitud, pues el denominador del error, el cual definimos como $e = z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$ será mayor, provocando

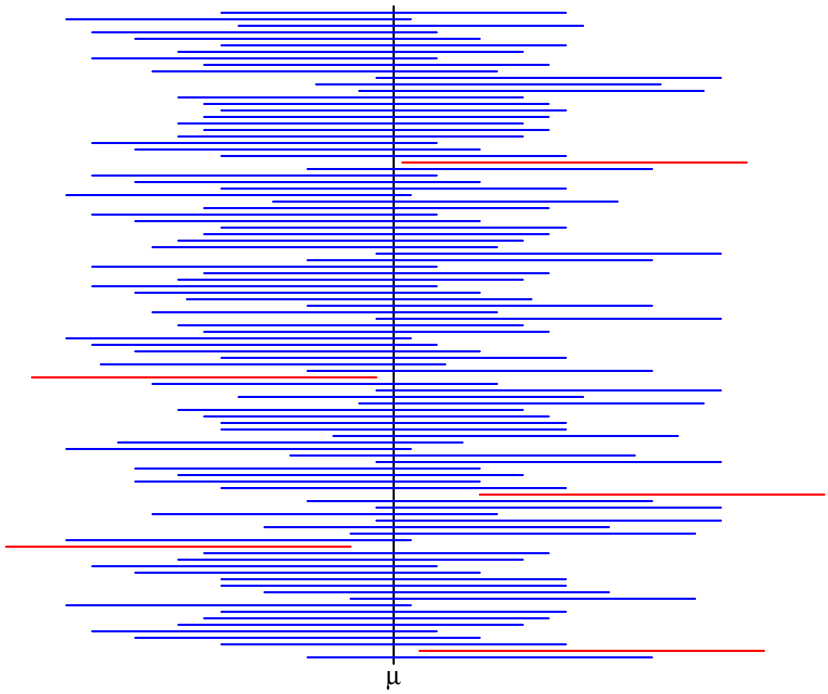


Figura 37: Intervalos de confianza para μ.

Nivel de Confianza $1-\alpha$	Percentil $z_{1-\alpha/2}$	Intervalo IC(μ)
$0{,}90$	$z_{0{,}95} = 1{,}645$	$\mu \in \left[\bar{X} - 1{,}645\frac{\sigma}{\sqrt{n}}, \bar{X} + 1{,}645\frac{\sigma}{\sqrt{n}}\right]$
$0{,}95$	$z_{0{,}975} = 1{,}960$	$\mu \in \left[\bar{X} - 1{,}960\frac{\sigma}{\sqrt{n}}, \bar{X} + 1{,}960\frac{\sigma}{\sqrt{n}}\right]$
$0{,}99$	$z_{0{,}995} = 2{,}576$	$\mu \in \left[\bar{X} - 2{,}576\frac{\sigma}{\sqrt{n}}, \bar{X} + 2{,}576\frac{\sigma}{\sqrt{n}}\right]$

Tabla 12: Intervalos de confianza para la media poblacional, para distintos niveles de confianza.

una disminución del error. Note que e corresponde a la diferencia máxima que podemos aceptar entre la estimación puntual del parámetro y el verdadero valor de dicho parámetro (es decir, lo que estamos dispuestos a sacrificar o equivocarnos).

Por último, si mantenemos fijo el nivel de confianza y el tamaño muestral, y aumentamos la desviación estándar, también aumentará la longitud del intervalo, pues esta vez el numerador del error será mayor. En otras palabras a mayor dispersión de los datos en la población, mayor será la longitud del intervalo.

Luego, podemos concluir que existen tres factores que afectan la longitud del intervalo de confianza:

- La elección del nivel de confianza $(1-\alpha)$.
- El tamaño muestral (n).
- La desviación estándar de la población (σ).

Note que en términos del error (e) el IC para μ puede ser escrito de la siguiente manera:

$$\mu \in \left[\bar{X}-e, \bar{X}+e\right],$$

como en estadística usualmente se requiere de un muestra para utilizar las herramientas o métodos disponibles, y los intervalos de confianza no son la excepción, debemos considerar una forma de determinar el tamaño muestral, para ello el error es fundamental, pues si despejamos adecuadamente a n, obtenemos:

$$n \geq \frac{z_{1-\alpha/2}^2 \sigma^2}{e^2}, \tag{4}$$

luego, como e es lo máximo que estamos dispuestos a equivocarnos, n será el tamaño muestral mínimo para que eso ocurra. Consideremos el siguiente ejemplo: Para estudiar el consumo de leche, en litros por persona al mes, se ha elegido una muestra de 50 personas cuyo consumo medio es de 22 litros. Si dicho consume sigue una distribución normal con desviación estándar de 6 litros, determine un intervalo de confianza para la media con un nivel de confianza del 95 $\%$.
Definiendo la variable en cuestión y traduciendo la información del problema tenemos:
Sea X_i : leche consumida por la i-ésima persona (en litros).
$X \sim N(\mu, 6^2)$, $n = 50$, $\bar{x} = 22$ litros, $1-\alpha = 0{,}95 \Longrightarrow z_{1-\alpha/2} = 1{,}96$.

Luego sustituyendo la información disponible, el IC será:

$$\begin{aligned} \mu \in &= \left[22 - 1{,}960\frac{6}{\sqrt{50}}; 22 + 1{,}960\frac{6}{\sqrt{50}}\right] \\ &= [22-1{,}66; 22+1{,}66] \\ &= [20{,}34; 23{,}66] \end{aligned} \tag{5}$$

Concluiremos diciendo que, con un 95 $\%$ de nivel de confianza el verdadero consumo de leche promedio esta entre $20{,}34$ y $23{,}66$ litros, donde el error es $1{,}66$ litros. Entonces si por ejemplo, deseamos disminuir el error de nuestra estimación máximo a un litro, ¿cuántas personas deberían pertenecer a la muestra?, la respuesta viene de sustituir esta nueva información en (4), de donde tendremos:

$$n \geq \frac{1{,}96^2 \times 6^2}{1^2} = 138{,}30 \approx 139 \text{ personas},$$

note que debemos aproximar el tamaño muestral al entero mayor, pues corresponde al mínimo de personas que deben componer la muestra para asegurar un error máximo de 1 litro en la estimación.
Note que a partir de un IC para la media podemos responder a preguntas como: ¿es posible concluir en base a esta muestra que el consumo promedio de leche de la población sea $20{,}8$ litros? La respuesta de acuerdo con (5) es sí, pues dicho valor pertenece al intervalo en cuestión. Para el intervalo construido con error máximo de 1 litro, la respuesta será no, pues el intervalo es $[21; 23]$ litros, y claramente $20{,}8$ no pertenece a él.

8.2.2. Didáctico

Por qué enseñar la distribución normal (y qué implica comprenderla)
La distribución normal constituye un *modelo* probabilístico fundamental para describir la variación continua en ciencias naturales y sociales, y especialmente para aproximar las distribuciones

muestrales de estadísticos mediante el *Teorema del Límite Central* (TCL). Comprenderla implica distinguir entre datos empíricos (p. ej., histogramas) y modelos teóricos (curva ideal), interpretar probabilidades como *áreas* bajo la curva y reconocer su carácter de idealización útil, más que de descripción universal de los datos observados.
Desde el punto de vista didáctico, se sugiere presentarla como una **familia de distribuciones** parametrizadas por la media (μ) y la desviación estándar (σ), enfatizando las transformaciones de *traslación* y *escala* que estos parámetros generan sobre la curva. Tal enfoque permite conectar la distribución normal con los procedimientos de **estandarización** (puntajes z y percentiles) y con la **toma de decisiones inferenciales** contextualizadas (Dinov et al., 2013; Salinas-Herrera & Salinas-Hernández, 2022).

8.2.3. Dificultades habituales (y por qué ocurren)

Conceptuales. Persisten confusiones entre *densidad* (altura de la curva) y *probabilidad* (área bajo la curva), la sobre–generalización del TCL (asumir normalidad de los *datos* en lugar de la distribución muestral) y la falta de diferenciación entre parámetros (μ, σ) y estadísticos $(\bar{x}, s)$. Estas confusiones dificultan el tránsito entre el razonamiento empírico y el razonamiento teórico-probabilístico.
Representacionales. La coordinación de múltiples registros (enunciado verbal $\rightarrow$ desigualdad $\rightarrow z \rightarrow$ sombreado $\rightarrow$ probabilidad $\rightarrow$ interpretación) constituye un obstáculo persistente, especialmente cuando los estudiantes no dominan la recta real ni el sentido de magnitud en números negativos (Delport, 2022).
Inferenciales y procedimentales. Se observan decisiones mecánicas apoyadas únicamente en pruebas de normalidad (p. ej., Shapiro–Wilk), sin diagnóstico gráfico ni análisis de robustez, así como el uso acrítico de reglas simplificadas (p. ej., $n \geq 30$). En estudiantes de primer año universitario, los bajos desempeños en problemas típicos de normal (cálculo de áreas, búsqueda de umbrales) revelan carencias en lectura matemática, modelización y verificación de resultados (Makwakwa et al., 2024).

8.2.4. Implicancias curriculares y de evaluación

Los hallazgos previos sugieren diseñar secuencias que integren datos y modelos desde etapas tempranas (p. ej., simulación del TCL) y evaluaciones que midan tanto la comprensión conceptual como la aplicación con sentido. En contextos de aprendizaje activo, se reportan mejoras significativas en el desempeño medio y en la homogeneización de la comprensión conceptual (Ridley et al., 2021). Ello refuerza el valor de estrategias de predicción, debate y verificación empírica.
La evaluación debería incluir: (a) traducciones semióticas completas (texto $\rightarrow$ desigualdad $\rightarrow z$ $\rightarrow$ área $\rightarrow$ interpretación); (b) juicio de *cuándo* la normal es un modelo razonable; y (c) el uso de representaciones gráficas (histograma, gráfico Q–Q) antes de aplicar pruebas formales.

8.2.5. Estrategias efectivas

Línea numérica como andamiaje. Las intervenciones que incorporan la línea numérica mejoran de manera significativa la resolución de tareas asociadas a la normal (delimitación de regiones, comprensión de z negativos), fortaleciendo el sentido numérico y el control de desigualdades (Delport, 2022).
Simulación y tecnología educativa. La simulación de muestreos repetidos (p. ej., con Fathom o entornos web interactivos) favorece el tránsito entre distribuciones empíricas y teóricas, reforzando la comprensión de la variabilidad y la aproximación normal (Salinas-Herrera & Salinas-Hernández, 2022). Recursos como el sistema SOCR permiten explorar la distribución normal (y

bivariada), manipular parámetros y visualizar probabilidades marginales o conjuntas, mostrando impactos positivos en la comprensión conceptual y la motivación estudiantil (Dinov et al., 2013). **Modelización con datos reales sencillos**. Actividades experimentales con medidas físicas simples (p. ej., masas de objetos cotidianos) facilitan la construcción de modelos normales ajustados, el análisis crítico de supuestos y la vinculación entre la curva teórica y fenómenos observables (Gomes & de Araújo Lima, 2012).

8.3. Orientaciones desde la implementación

El problema del SERNAC sobre el peso de las bolsas de azúcar permite articular de manera natural los conceptos de *estimación puntual*, *intervalo de confianza*, *variabilidad muestral* y *razonamiento inferencial*. Su implementación debe orientarse a que los estudiantes comprendan que el objetivo no es sólo aplicar una fórmula, sino argumentar sobre el grado de evidencia que la muestra entrega respecto al cumplimiento de la normativa de etiquetado (1 kg).

1) **Contextualización y sentido del problema.** Inicie la sesión planteando una situación real: el control de calidad de productos de consumo masivo. Pregunte a los estudiantes qué significaría que una fábrica declare "1 kg" y cómo podrían verificarlo. Esto activa conocimientos previos sobre variabilidad y muestreo, además de despertar interés por el contexto regulatorio.
2) **Exploración de la muestra.** Presente los datos o un resumen muestral ($n = 850$, $\bar{x} = 996\,\text{g}$) y discuta qué información falta (la variabilidad). Proponga distintos escenarios para s (p. ej., 10 g, 30 g, 60 g) y pida a los estudiantes predecir, sin calcular aún, si el valor de $1\,000\,\text{g}$ estaría o no dentro del intervalo. Esto promueve el pensamiento estimativo y la anticipación inferencial antes de proceder al cálculo formal.
3) **Construcción guiada del intervalo.** A partir del caso base, conduzca la formulación del intervalo:

$$\mu \in \bar{x} \pm z_{0{,}975}\frac{s}{\sqrt{n}}.$$

   Utilice representaciones gráficas (línea numérica o recta real) para visualizar la media muestral y los límites inferior y superior. Destaque que el margen de error depende de la *variabilidad* (s) y del *tamaño muestral* (n).
4) **Interpretación y argumentación.** Guíe la discusión hacia el sentido del intervalo: no se trata de una probabilidad de que μ esté dentro del rango, sino del nivel de confianza asociado al procedimiento. Pregunte: *¿Qué implica si $1\,000\,g$ no está dentro del intervalo? ¿Y si lo está justo en el límite?* Esto fomenta la argumentación y el juicio crítico sobre la evidencia empírica y las decisiones regulatorias.
5) **Conexión con la variabilidad y la simulación.** Si dispone de software (R, GeoGebra, o applets de SOCR), simule distintas muestras de tamaño 850 con media 996 y distintas desviaciones estándar, graficando los intervalos obtenidos. Esta simulación permite visualizar la frecuencia con que los intervalos capturan el valor verdadero y comprender empíricamente el significado del 95 % de confianza (Chance et al., 2002; Dinov et al., 2013).
6) **Extensión y transferencia.** Proponga que los estudiantes redacten un informe técnico dirigido al SERNAC, indicando si el proceso de producción cumple con la normativa y justificando la decisión con base en el intervalo obtenido. De esta forma, se fortalece la conexión entre el lenguaje estadístico y la comunicación profesional basada en evidencia (Ridley et al., 2021).

Estas orientaciones buscan que el problema se convierta en una *tarea de indagación* y no en un mero ejercicio algorítmico, promoviendo el razonamiento inferencial, la interpretación contextual y la valoración de la incertidumbre estadística como parte del pensamiento científico.

Bibliografía

Ainsworth, S. (1999). The functions of multiple representations. *Computers & Education*, *33*(2-3), 131-152. https://doi.org/10.1016/S0360-1315(99)00029-9

American Statistical Association & National Council of Teachers of Mathematics. (2020). *Pre-K–12 Guidelines for Assessment and Instruction in Statistics Education II (GAISE II): A Framework for Statistics and Data Science Education*.

Appova, A., & Taylor, C. E. (2020). Providing opportunities to develop prospective teachers' pedagogical content knowledge. *The Mathematics Enthusiast*, *17*(2-3), 673-724. https://scholarworks.umt.edu/tme/vol17/iss2/12/

Aridor, K., & Ben-Zvi, D. (in press). Informal inferential reasoning: Signal and noise in students' statistical thinking. *Statistics Education Research Journal*.

Bakker, A., & Derry, J. (2011). Philosophy of statistics education. En P. S. Bandyopadhyay & M. Forster (Eds.), *Philosophy of Statistics* (pp. 497-521). Elsevier.

Ball, D. L., Thames, M. H., & Phelps, G. (2008). Content knowledge for teaching: What makes it special? *Journal of Teacher Education*, *59*(5), 389-407. https://doi.org/10.1177/0022487108324554

Batanero, C., Chernoff, E. J., & Engel, J. (2018). *Teaching and Learning of Statistics: International Perspectives*. Springer.

Ben-Zvi, D. (2004). Reasoning about variability in comparing distributions. *Statistics Education Research Journal*, *3*(2), 42-63. https://doi.org/10.52041/serj.v3i2.547

Ben-Zvi, D. (2006). Advancing reasoning about variation: Insights from informal inferential reasoning in a large data set context. *International Journal of Computers for Mathematical Learning*, *11*(1), 95-134.

Ben-Zvi, D., & Arcavi, A. (2001). Junior high school students' construction of global views of data and data representations. *Educational Studies in Mathematics*, *45*(1), 35-65.

Ben-Zvi, D., Makar, K., Garfield, J., & Aridor, K. (2012). Learning to reason about uncertainty: Developing informal inferential reasoning. *Proceedings of the IASE Roundtable Conference*.

Burgess, T. (2009). Teacher knowledge and statistics: What types of knowledge are used in the primary classroom? *The Mathematics Enthusiast*, *6*(1-2), 3-24. https://doi.org/10.54870/1551-3440.1130

Catalán, D., Pérez, B., Prieto, C., & Rupin, P. (2017). Texto del Estudiante. Matemática 8º Básico. *Santiago: SM Editorial*.

Chance, B. L., Garfield, J. B., & delMas, R. C. (2002). Use of technology in an introductory statistics course: A survey of college instructors. *The American Statistician*, *56*(4), 283-294.

Cobb, G. W. (2007). The thinking of statisticians and the thinking of students. *Proceedings of the IASE Satellite Conference*.

Cobb, G. W., & Moore, D. S. (1997). Mathematics, statistics, and teaching. *American Mathematical Monthly*, *104*(9), 801-823.

Croucher, J. S. (2013). *Introductory Statistics: A Conceptual Approach Using R*. McGraw-Hill Education.

Curcio, F. R. (1987). Comprehension of mathematical relationships expressed in graphs. *Journal for research in mathematics education*, *18*(5), 382-393.

delMas, R. C., Garfield, J., Ooms, A., & Chance, B. (2007). Assessing students' conceptual understanding after a first course in statistics (CAOS). *Statistics Education Research Journal*, *6*(2), 28-58.

Delport, L. (2022). Using the number line to improve students' understanding of negative numbers and normal distributions. *African Journal of Research in Mathematics, Science and Technology Education*, *26*(1), 67-83. https://doi.org/10.1080/18117295.2022.2030014

Díaz-Levicoy, D., Morales, M., Vásquez, C., & López-Martín, M. (2015). Niveles de lectura e interpretación de gráficos estadísticos en estudiantes de educación básica. *Revista Latinoamericana de Investigación en Matemática Educativa*, *18*(2), 229-252.

Dinov, I. D., Kamino, S., Bhakhrani, B., & Christou, N. (2013). Normal distribution: history, properties, and applications. *Statistics Online Computational Resource (SOCR) Educational Materials*. https://www.socr.umich.edu/

Fielding-Wells, J. (2010). Linking data and context: Developing students' informal inferential reasoning. *Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8)*.

Friel, S. N., Curcio, F. R., & Bright, G. W. (2001). Making sense of graphs: Critical factors influencing comprehension and instructional implications. *Journal for Research in mathematics Education*, *32*(2), 124-158.

Fuson, K. C., Kalchman, M., & Bransford, J. D. (2005). Mathematical understanding: An introduction. En M. S. Donovan & J. D. Bransford (Eds.), *How students learn: Mathematics in the classroom* (pp. 217-256). National Academies Press.

Garfield, J., & Ben-Zvi, D. (2008). *Developing students' statistical reasoning: Connecting research and teaching practice*. Springer. https://doi.org/10.1007/978-1-4020-8383-9

Gomes, A. M., & de Araújo Lima, T. (2012). Teaching normal distribution using simple physical measures. *Teaching Statistics*, *34*(3), 108-111. https://doi.org/10.1111/j.1467-9639.2011.00465.x

Groth, R. E. (2007). Toward a conceptualization of statistical knowledge for teaching. *Journal for Research in Mathematics Education*, *38*(5), 427-437. https://pubs.nctm.org/view/journals/jrme/38/5/article-p427.xml

Hancock, C., Kaput, J. J., & Goldsmith, L. T. (1992). Teaching statistics: More than computation. *Mathematics Teacher*, *85*(1), 46-48.

Harradine, A., Batanero, C., & Rossman, A. (2011). Developing teachers' statistical reasoning: What can we learn from research? *Proceedings of the 58th World Statistics Congress of the International Statistical Institute*.

Hirai, Y. (1989). Some remarks on class interval of histograms., *82*(1), 113-117.

Johnson Lachuk, A., Gísladóttir, K. R., & DeGraff, T. (2019). Using collaborative inquiry to prepare preservice teacher candidates who have integrity and trustworthiness. *Teaching and Teacher Education*, *78*, 75-84. https://doi.org/10.1016/j.tate.2018.11.003

Kader, G. D., & Perry, M. (2007). Variability for categorical variables. *Journal of Statistics Education*, *15*(2).

Kaput, J. J. (1989). Linking representations in the symbol systems of algebra. En S. Wagner & C. Kieran (Eds.), *Research issues in the learning and teaching of algebra* (pp. 167-194). Lawrence Erlbaum Associates.

Konold, C., Finzer, W., & Kreetong, K. (2015). Modeling as a core component of structuring data. *Statistics Education Research Journal*, *14*(2), 7-28.

Konold, C., & Pollatsek, A. (2002). Data analysis as the search for signals in noisy processes. *Journal for Research in Mathematics Education*, *33*(4), 259-289.

Kosslyn, S. M. (1985). Graphics and human information processing: A review of five books. *Journal of the American Statistical Association*, *80*(391), 499-512.

Lesh, R., Post, T., & Behr, M. (1987). Representations and translations among representations in mathematics learning and problem solving. En C. Janvier (Ed.), *Problems of representation in the teaching and learning of mathematics* (pp. 33-40). Lawrence Erlbaum Associates.

Makar, K., & Confrey, J. (2004). Secondary teachers' statistical reasoning in comparing two groups. En D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 353-373). Springer. https://doi.org/10.1007/1-4020-2278-6_15

Makar, K., & Rubin, A. (2009). Teachers' informal inferential reasoning. En D. Ben-Zvi & J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning and Thinking* (pp. 277-294). Springer.

Makwakwa, N., Mogari, D., & Ogbonnaya, U. I. (2024). Difficulties experienced by undergraduate students in solving problems involving the normal distribution. *International Journal of Mathematical Education in Science and Technology*. https://doi.org/10.1080/0020739X.2024.2371811

McLeod, S. (2019). *Empirical Rule in Statistics: Definition, Formula & Example* [Consultado el 28 de octubre de 2025]. https://www.simplypsychology.org/empirical-rule.html

Ministerio de Desarrollo Social y Familia. (2017). Informe de Desarrollo Social 2017 [Consultado el 18 de junio de 2025]. https://www.desarrollosocialyfamilia.gob.cl/pdf/upload/IDS2017.pdf

Ministerio de Educación de Chile. (2022). Datos Abiertos Mineduc [Consultado el 19 de junio de 2025]. https://datosabiertos.mineduc.cl/

Ministerio de Educación de Chile (Mineduc). (2015). *Ajuste Curricular de Matemática para Educación Media*. Mineduc.

Ochoa, J. (2019). Comprensión gráfica y niveles de interpretación de datos estadísticos en educación secundaria. *Revista Latinoamericana de Investigación en Matemática Educativa*, *22*(3), 241-266.

Perry, M., & Kader, G. D. (2005). Variation as unalikeability. *Teaching Statistics*, *27*(2), 58-60.

Pfannkuch, M. (2006). Comparing box plot distributions: A teacher's reasoning. *Statistics Education Research Journal*, *5*(2), 27-45.

Reading, C., & Reid, J. (2006). An emerging hierarchy of reasoning about distribution: From a variation perspective. *Statistics Education Research Journal*, *5*(2), 46-68.

Ridley, K., Ngnepieba, P., & de Silva, D. (2021). Active learning strategies in introductory statistics: Effects on performance and conceptual understanding. *Journal of Statistics Education*, *29*(2), 143-159. https://doi.org/10.1080/10691898.2021.1917632

Rossman, A. (2008). Developing a concept of statistical inference. *Proceedings of the IASE Roundtable Conference on International Perspectives on the Teaching and Learning of Statistics*, 1-8.

Rubin, A., Hammerman, J. K., & Konold, C. (2006). Representing, comparing, and interpreting data with variability. *Statistics Education Research Journal*, *5*(2), 52-73.

Ruz, F., Molina-Portillo, E., & Contreras, J. M. (2020). Actitudes hacia la estadística descriptiva y su enseñanza en futuros profesores. *Cadernos de Pesquisa*, *50*, 964-980.

Salinas-Herrera, J., & Salinas-Hernández, J. (2022). Didactic approaches to teaching the central limit theorem through simulation and visualization. *Revista Latinoamericana de Investigación en Matemática Educativa*, *25*(3), 311-333. https://doi.org/10.12802/relime.22.2534

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher*, *15*(2), 4-14. https://doi.org/10.3102/0013189X015002004

Sjølie, E., Strømme, A., & Boks-Vlemmix, J. (2021). Team-skills training and real-time facilitation as a means for developing student teachers' learning of collaboration. *Teaching and Teacher Education*, *107*, 103477. https://doi.org/10.1016/j.tate.2021.103477

Sturges, H. A. (1926). The choice of a class interval. *Journal of the american statistical association*, *21*(153), 65-66.

Tukey, J. W. (1980). We need both exploratory and confirmatory. En P. R. Krishnaiah (Ed.), *The Role of Data Analysis in the Statistical Process* (pp. 23-41). Elsevier.

Tukey, J. W., et al. (1977). *Exploratory data analysis* (Vol. 2). Springer.

Vásquez, C., & Alsina, A. (2014). Enseñanza de la probabilidad en educación primaria. Un desafío para la formación inicial y continua del profesorado. *Números*, *85*(1), 5-23.

Vásquez, C., & Alsina, Á. (2021). Analysing probability teaching practices in primary education: what tasks do teachers implement? *Mathematics*, *9*(19), 2493.

Wainer, H. (1992). Understanding graphs and tables. *Educational researcher*, *21*(1), 14-23.

Wild, C. J., & Pfannkuch, M. (1999a). Statistical thinking in empirical enquiry. *International Statistical Review*, *67*(3), 223-265. https://doi.org/10.1111/j.1751-5823.1999.tb00442.x

Wild, C. J., & Pfannkuch, M. (1999b). Statistical thinking in empirical enquiry. *International statistical review*, *67*(3), 223-248.

Wilensky, U. (1995). Learning probability through building computational models. *Journal of Computer Assisted Learning*, *11*(4), 195-203.

Wilensky, U. (1997). Epistemological anxiety and computational experiments: A new approach to understanding probability learning. *Proceedings of the First International Conference on Computers and Mathematics Learning*, 237-265.

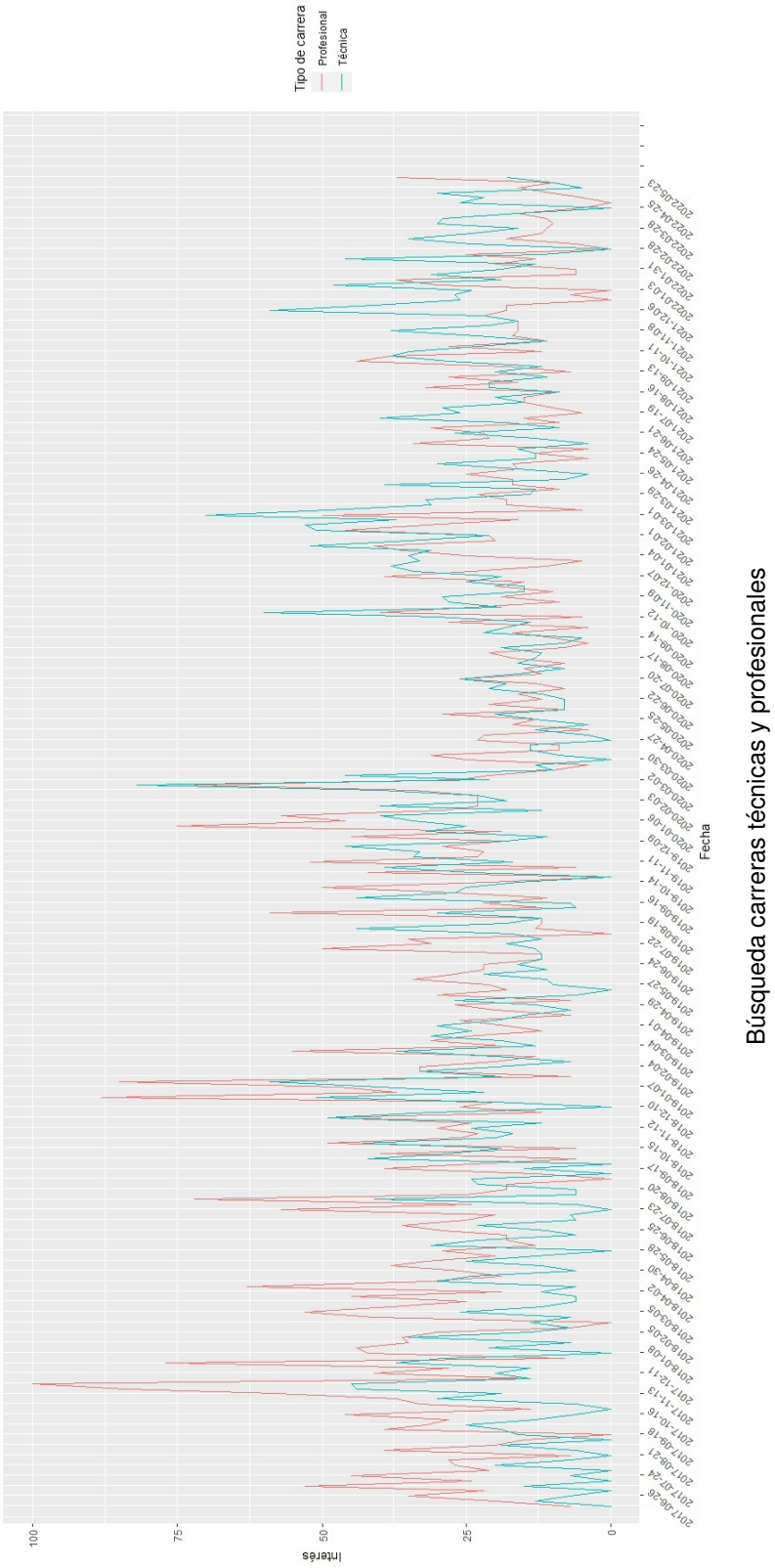
Zieffler, A., Garfield, J., delMas, R., & Reading, C. (2008). Developing a modern curriculum for introductory statistics: Focus on statistical literacy, reasoning, and thinking. *Proceedings of the IASE Roundtable Conference*.

9. Anexos

Los anexos que se presentan a continuación reúnen los problemas trabajados a lo largo del libro. Su propósito es ofrecer a las y los docentes una recopilación accesible y organizada de estas situaciones, de modo que puedan ser fácilmente reproducidas y utilizadas en el aula.

Problema 1: ¿Una carrera técnica o una profesional?

Observa el gráfico y explica por qué se producen las variaciones búsqueda en el rango de tiempo.



Simplificación

- ¿Por qué en los meses de noviembre a enero aumenta la búsqueda de carreras profesionales y técnicas?
- ¿Por qué en los meses de julio y agosto la búsqueda de carreras técnicas aumenta pero no las profesionales?
- En el primer semestre del año 2020, ¿por qué la búsqueda de ambos tipos de carrera descendieron?

Extensión

Observe la siguiente información sobre la matrícula del primer año según el tipo de institución de educación superior.

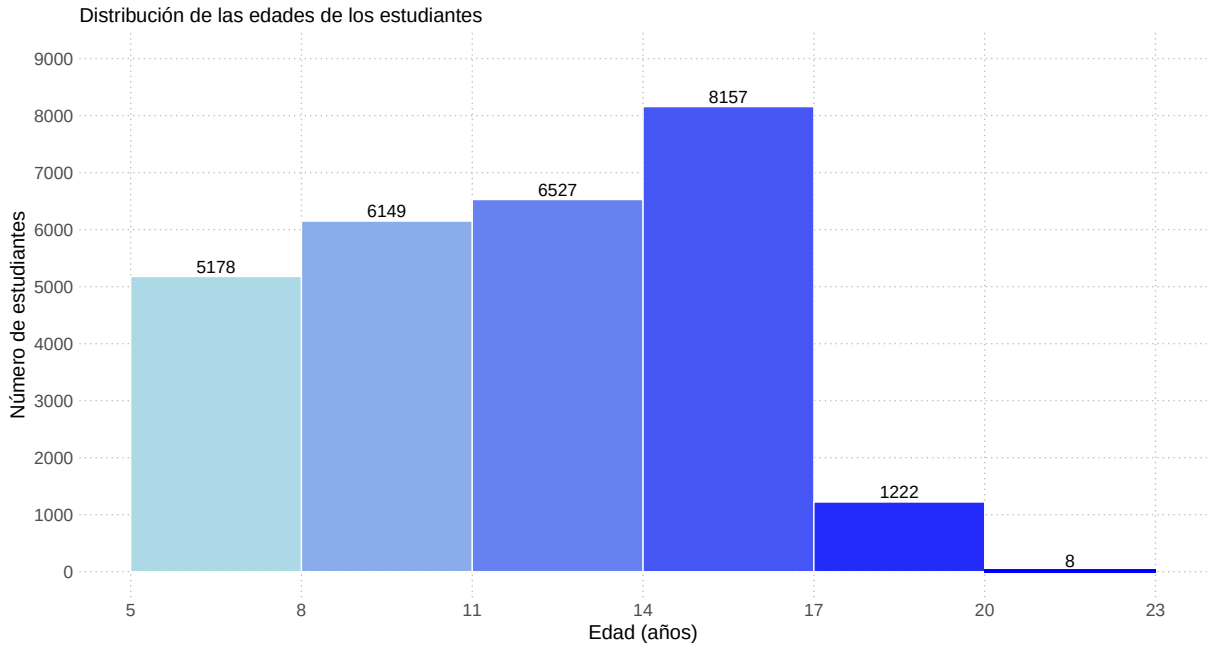
Tipo de IES	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
Universidad	154.931	153.107	149.617	153.727	162.132	166.016	167.482	160.826	149.966	147.331
Instituto Profesional	108.381	124.900	126.907	124.344	125.389	120.430	123.536	126.497	114.456	127.131
Centro de Formación Técnica	60.870	63.623	65.610	62.551	59.136	59.300	59.768	61.744	56.253	58.754

Matrículas de Primer Año, periodo 2012-2021. Fuente: Elaboración propia a partir de ÍNDICES Matrículas de Educación Superior 2005-2021, CNED.

Con la información de la tabla, evalúe las conclusiones que se pueden obtener del gráfico de línea.

Problema 2: De la montaña a la caja.

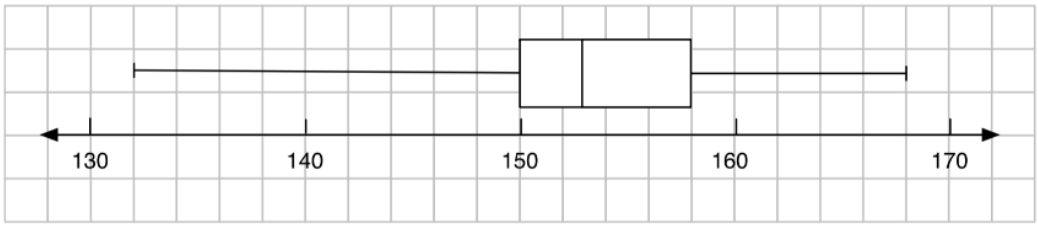
El siguiente histograma muestra la altura en centímetros de las niñas de un curso.



Construya el gráfico de cajas asociado.

Una simplificación

El siguiente gráfico de caja muestra la distribución de las alturas de un grupo de niñas.



Construya el histograma de la distribución de las alturas de las niñas.

Una extensión

Los siguientes histogramas muestran la distribución de puntajes asociados a una evaluación. Construya para cada grupo el gráfico de cajas asociado.

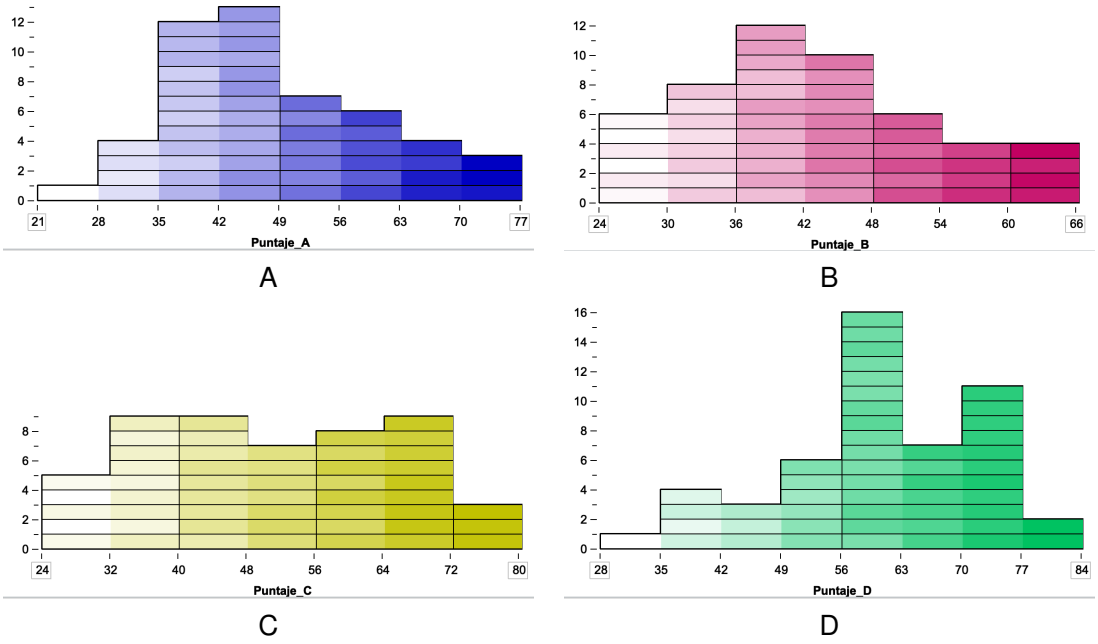


Tabla 13: Distribución de puntajes de 4 evaluaciones.

Problema 3: Suavizando hacia la caja.

El gráfico de distribución suavizado muestra el número de gigabytes que gasta un grupo de estudiantes en un mes de celular.

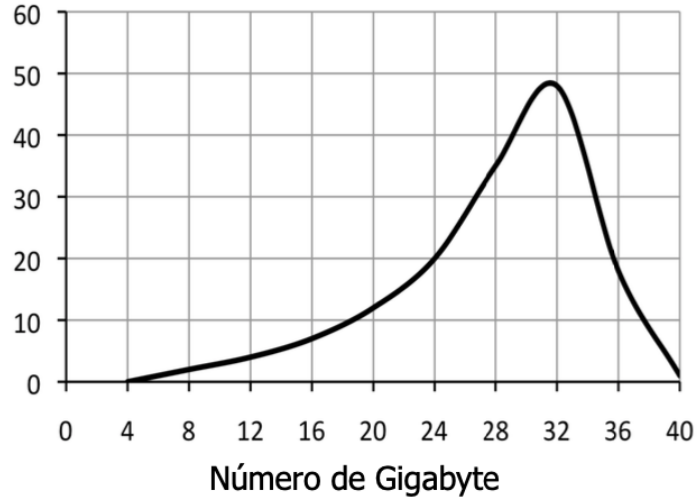
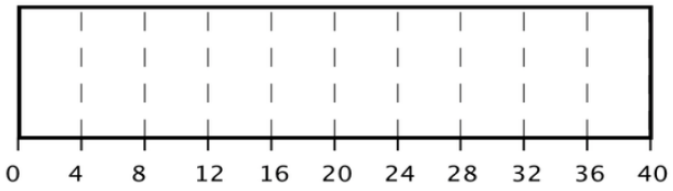


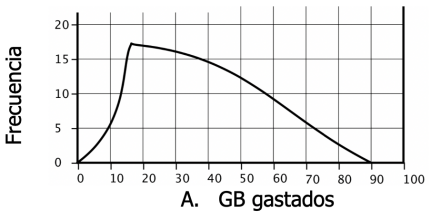
Gráfico de caja

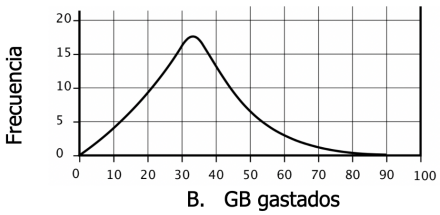


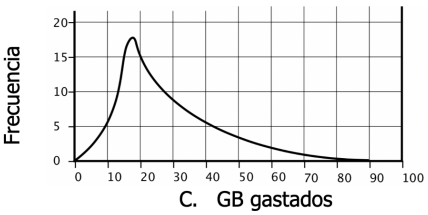
Construye el posible gráfico de caja asociado.

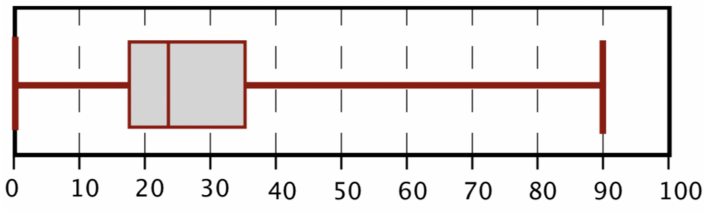
Una simplificación

Los gráficos de distribución suavizado muestran el número de gigabyte que gasta un grupo de estudiantes en un mes de celular. ¿Cuál de los gráficos corresponde al gráfico de caja?





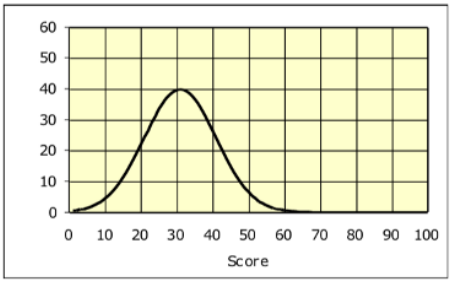




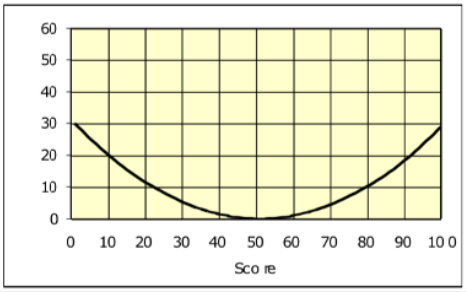
Una extensión

Relaciona cada distribución suave con el gráfico de caja correspondiente

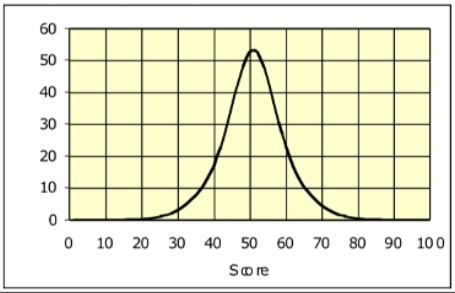
Frequency Graph A



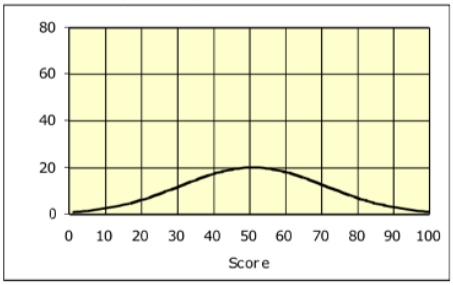
Frequency Graph B



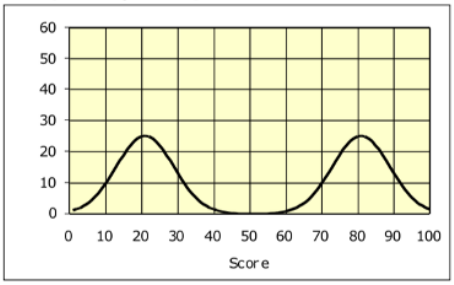
Frequency Graph C



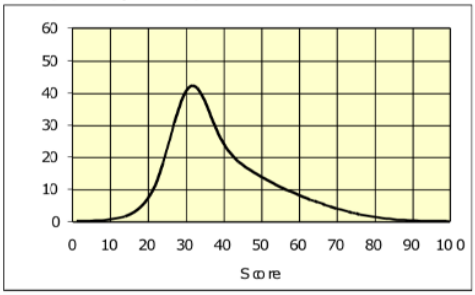
Frequency Graph D



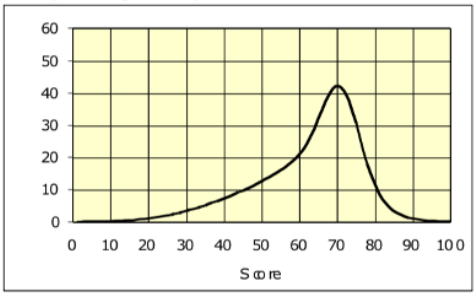
Frequency Graph E



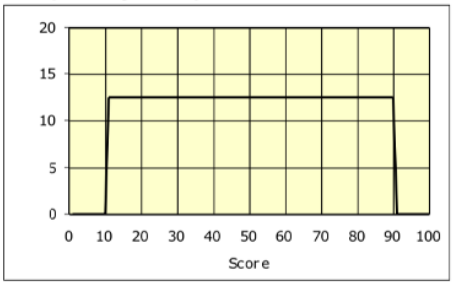
Frequency Graph F



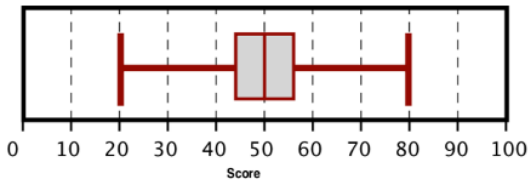
Frequency Graph G



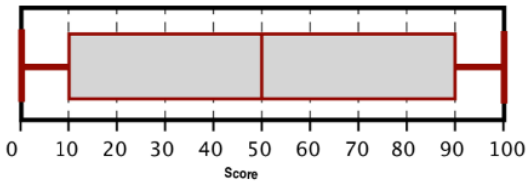
Frequency Graph H



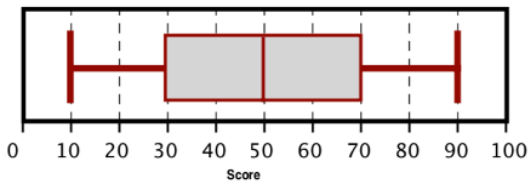
Box Plot 1



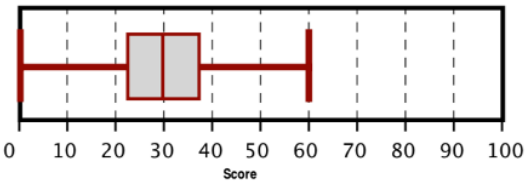
Box Plot 2



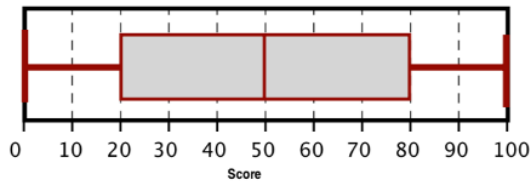
Box Plot 3



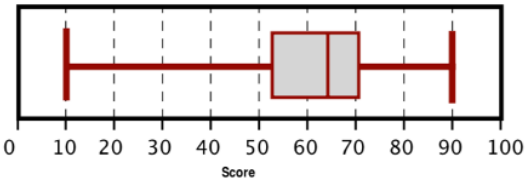
Box Plot 4



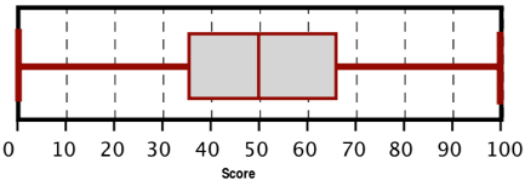
Box Plot 5



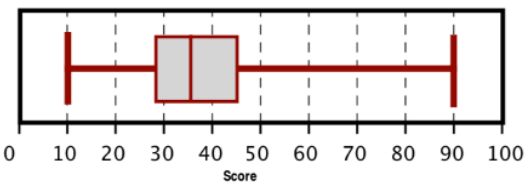
Box Plot 6



Box Plot 7

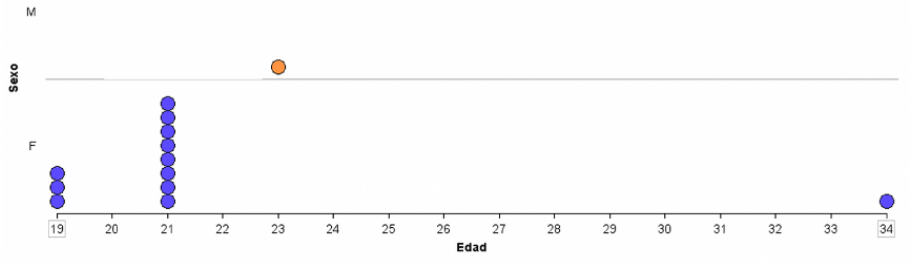


Box Plot 8



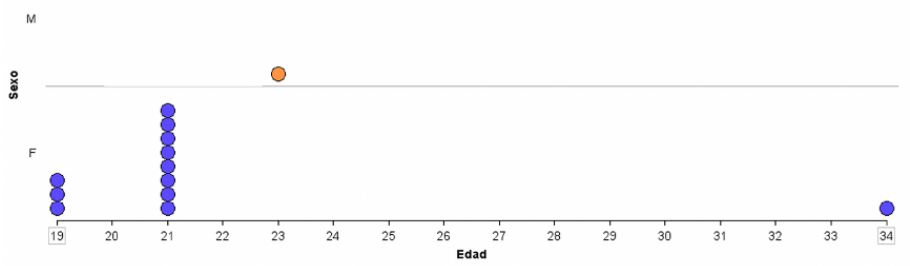
Problema 4: La Edad Media

Los 3 hombres de un curso tienen una edad promedio de 22 años y el curso entero, de 36 estudiantes, tienen una edad promedio de 24 años. Representa los puntos de las personas que faltan en el gráfico de distribución.



Una simplificación

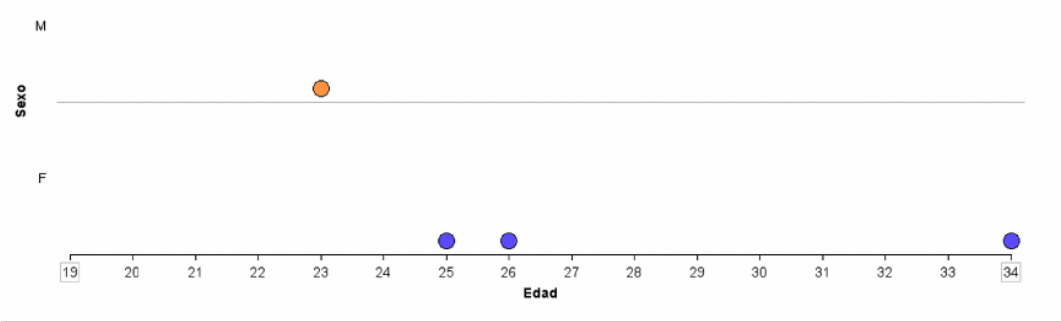
Los 3 hombres de un curso tienen una edad promedio de 22 años y el curso entero, de 36 estudiantes, tienen una edad promedio de 24 años. Representa solo los puntos de los hombres que faltan.



Una extensión

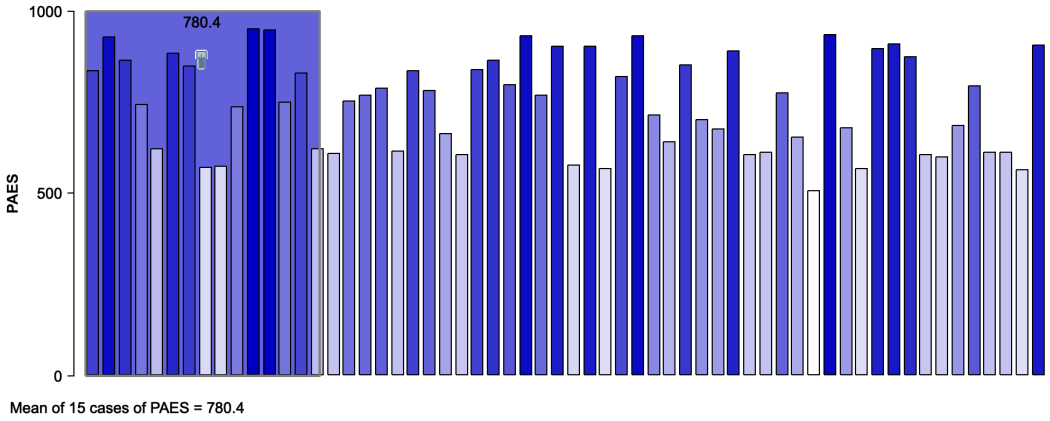
La Edad Media

La edad promedio de un curso es igual a 24 años. Si se compone de hombres y mujeres. Completa los puntos que faltan en la distribución y determina las edades promedio de cada sexo.

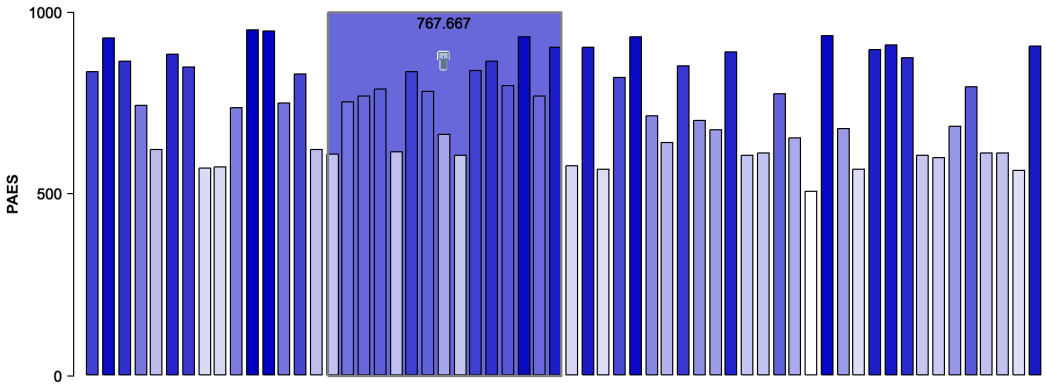


Problema 5: La PAES

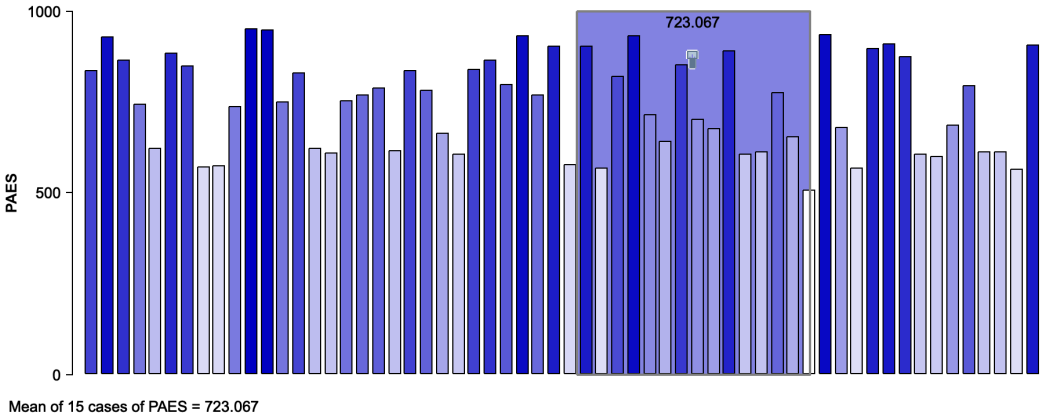
Las siguientes imágenes visualizan la distribución de puntajes de dos cursos de estudiantes de cuarto medio en la prueba PAES. Cada cajón muestra la media muestral de 15 estudiantes.



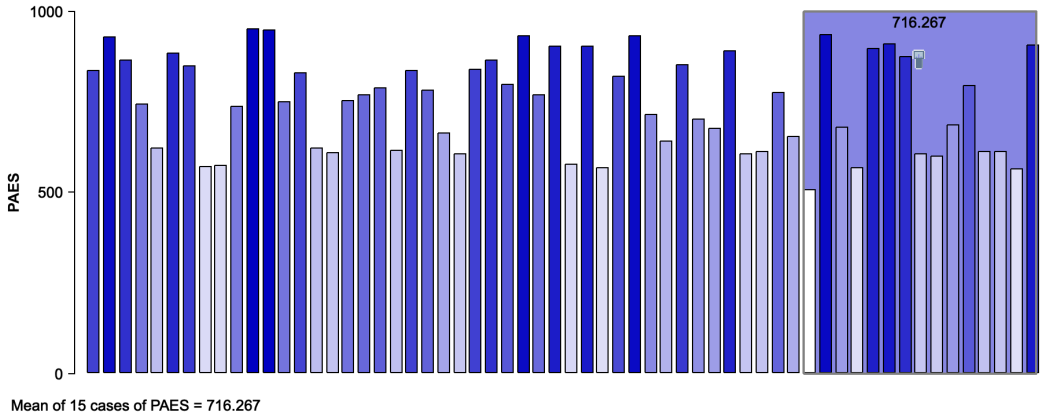
Mean of 15 cases of PAES = 780.4



Mean of 15 cases of PAES = 767.667



Mean of 15 cases of PAES = 723.067

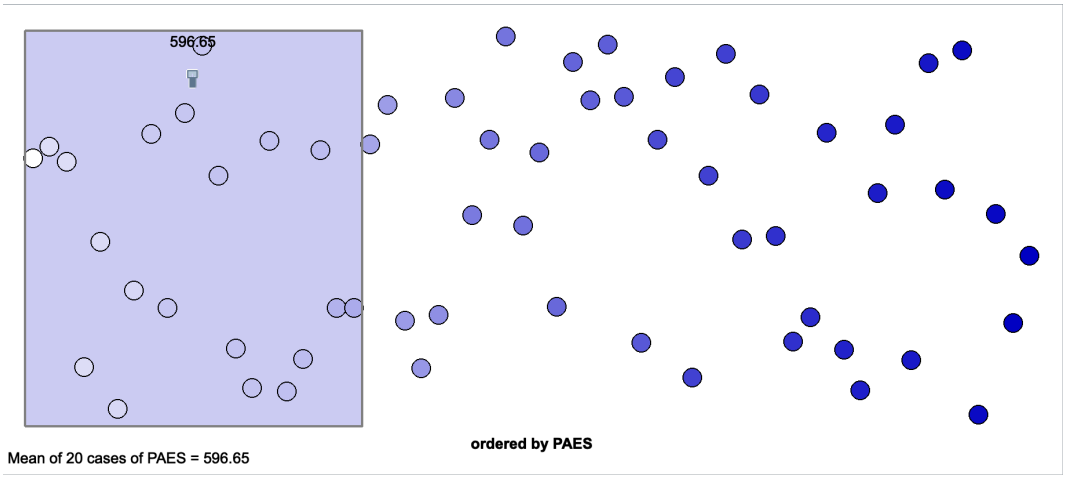


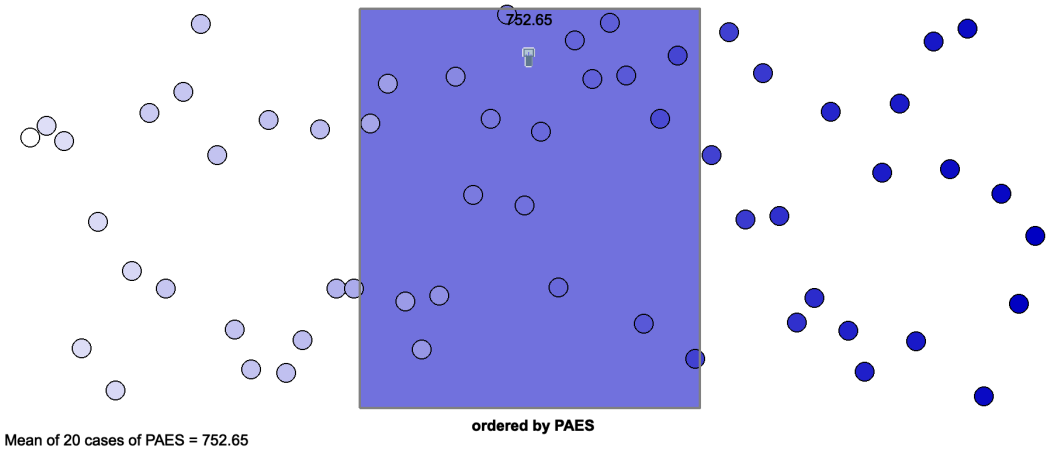
Mean of 15 cases of PAES = 716.267

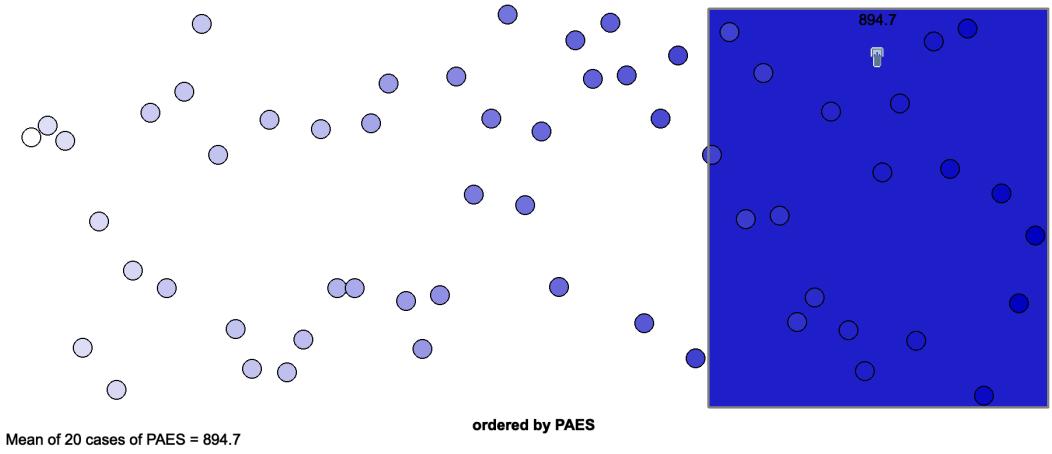
¿Cuál es la media de los dos cursos de estudiantes?

Una simplificación

Las siguientes imágenes visualizan la distribución de puntajes de dos cursos de estudiantes de cuarto medio en la prueba PAES. Cada cajón muestra la media muestral de 20 estudiantes.



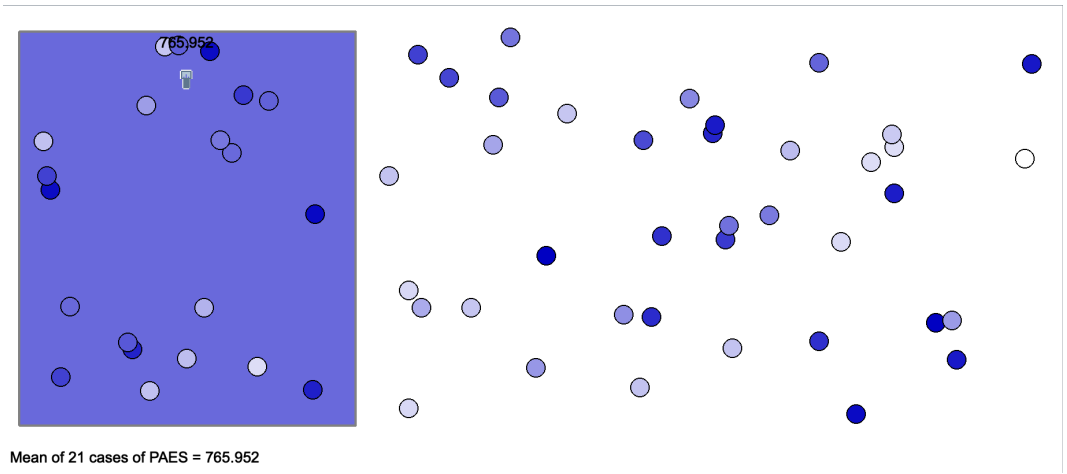




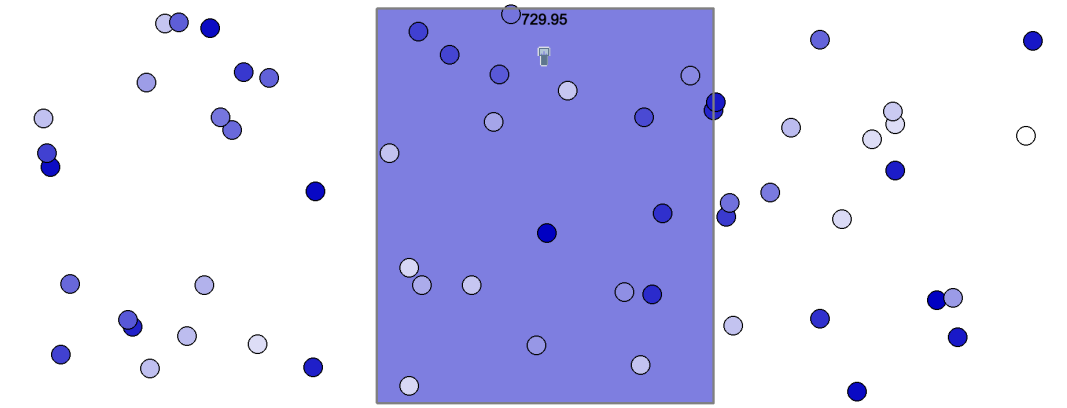
¿Cuál es la media de los dos cursos de estudiantes?

Una extensión

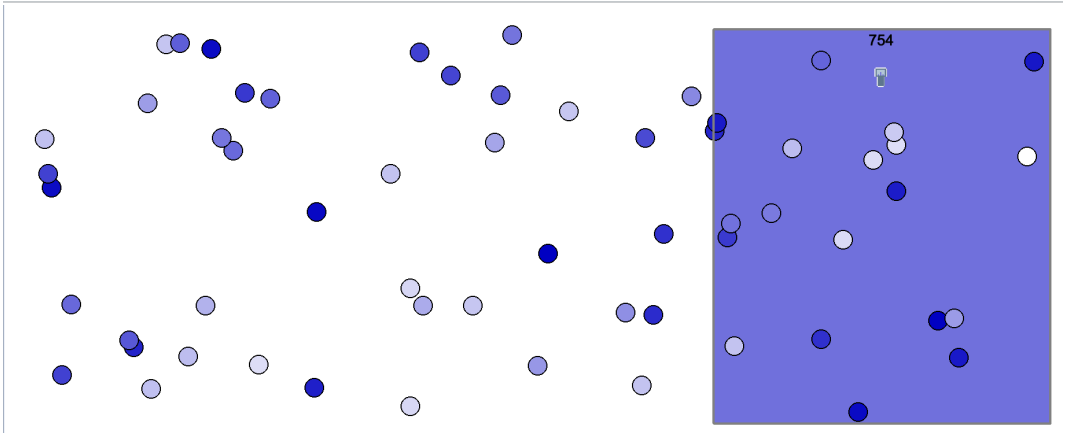
Las siguientes imágenes visualizan la distribución de puntajes de dos cursos de estudiantes de cuarto medio en la prueba PAES. Cada cajón muestra la media muestral de un grupo de estudiantes.



Mean of 21 cases of PAES = 765.952



Mean of 20 cases of PAES = 729.95

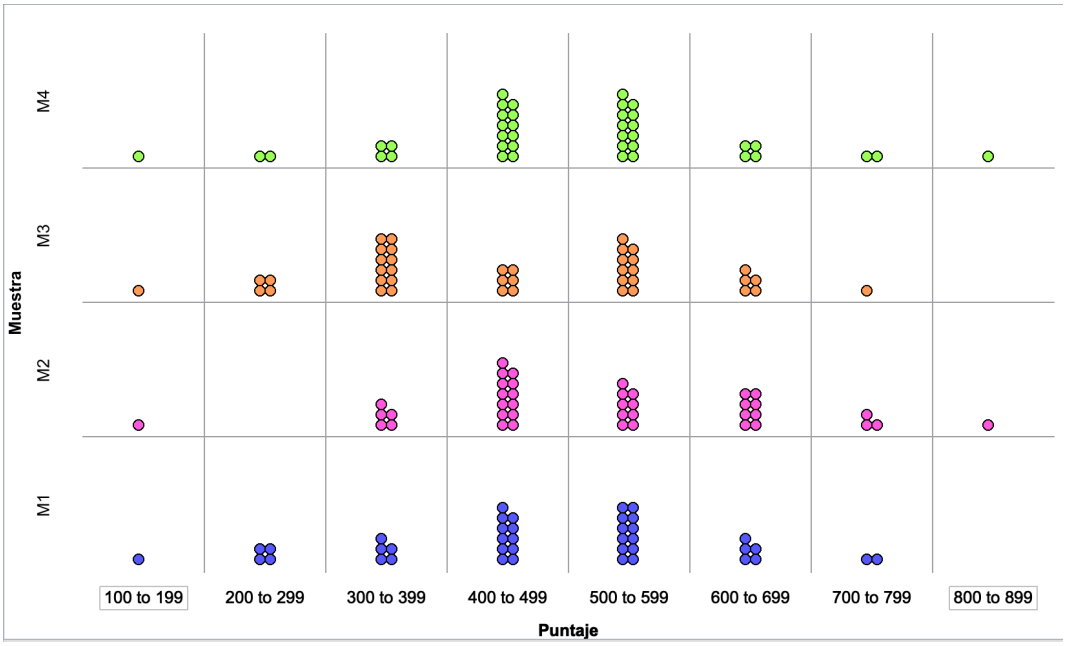


Mean of 20 cases of PAES = 754

¿Cuál es la media de los dos cursos de estudiantes?

Problema 6: El parámetro

La siguiente figura representa 4 muestras aleatorias que pertenecen a una población de puntajes PAES que tiene una desviación de 150 puntos. ¿Cuál puede ser la media de la población?

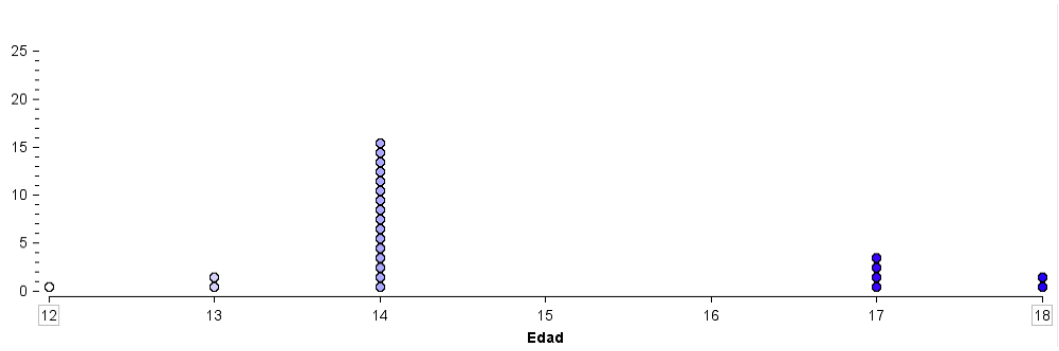


Una simplificación Las siguientes 4 muestras pertenecen a una población de puntajes PAES cuya distribución es simétrica. ¿Cuál puede ser la media de la población de puntajes?

M1	391	269	396	518	295	588	543	364	534	612
	344	241	596	271	500	538	585	528	390	648
	659	468	486	487	716	669	625	457	556	560
M2	761	632	208	710	492	579	593	485	489	653
	731	484	577	532	472	482	652	470	194	471
	607	649	857	600	531	168	904	428	856	556
M3	581	592	592	246	555	645	691	466	452	723
	622	636	500	323	302	411	620	206	217	402
	250	434	569	257	542	782	499	458	571	458
M4	124	265	488	531	542	623	471	682	362	319
	559	363	633	550	474	623	558	433	535	597
	553	401	628	673	541	522	489	824	541	476

Una extensión

La edad de los estudiantes de dos cursos se distribuye normal con media 15 años. Si entre los dos cursos suman 70 estudiantes, ¿cuál es la edad de los estudiantes que no están representados en el gráfico de puntos?



Problema 7: La línea de Maradona

El Sernac está inspeccionando las bolsas de azúcar producidas por la firma ACME-Maradona, que supuestamente traen un kilo cada una. Se toma una muestra aleatoria simple de 850 bolsas de una partida. El peso promedio resulta ser 996 g. ¿Cuál es el intervalo de confianza para el peso promedio real de las bolsas?

Una simplificación

El Sernac está inspeccionando las bolsas de azúcar producidas por la firma ACME-Maradoma, que supuestamente traen un kilo cada una. Se toma una muestra aleatoria simple de 850 bolsas de una partida. El peso promedio resulta ser 996 g, con una desviación típica de 50 g. ¿En que casos intervalo de confianza para el peso promedio real de las bolsas contiene el valor anunciado de 1000g?

Una extensión

El Sernac está inspeccionando las bolsas de azúcar producidas por la firma ACME-Maradoma, que supuestamente traen un kilo cada una. Se toma una muestra aleatoria simple de bolsas de una partida. El peso promedio resulta ser 996 g. ¿Cuál es el intervalo de confianza para el peso promedio real de las bolsas al 90 %, al 95 % y al 99 %?