TOUR GUÍADO POR LA ESTADÍSTICA BÁSICA

CONCEPTOS, MAPAS, VIDEOS Y MÁS

JOSÉ MANUEL MAGALLANES

COLABORADORES

LUIS ALBERTO MAS CASTILLO

NOAM VALENTÍN LÓPEZ VILLANES

MARIELA DEL PILAR MOSQUEIRA CABRERA

LORENA LÉVANO GAVIDIA

TOUR GUÍADO POR LA ESTADÍSTICA BÁSICA

CONCEPTOS, MAPAS, VIDEOS Y MÁS

JOSÉ MANUEL MAGALLANES

Pontificia Universidad Católica del Perú

COLABORADORES

LUIS ALBERTO MAS CASTILLO

NOAM VALENTÍN LÓPEZ VILLANES

MARIELA DEL PILAR MOSQUEIRA CABRERA

LORENA LÉVANO GAVIDIA

PONTIFICIA
UNIVERSIDAD
CATÓLICA
DEL PERÚ

Lima, Perú

Enero 2012

Tour guiado por la estadística

conceptos, vídeos, mapas y más

Pontificia Universidad Católica del Perú
Av. Universitaria 1801, San Miguel, Lima 32, Perú
Teléfono (511) 626-2000
www.pucp.edu.pe

Primera edición :
Lima, febrero de 2012
ISBN: XXXXXXXXX 9786124057557

Hecho el Depósito Legal en la Biblioteca Nacional del Perú
N° XXXXXX

Diseño de carátula: REP S.A. C.
Diagramación: REP S.A. C.
Impreso en Perú por:
REP SAC
Cervantes 485-502, San Isidro
Lima 27,Perú
Teléfonos: 421-5712 / 999-658531
jcandiotti@revistasespecializadas.com

Agradecimientos

Al Departamento de Ciencias Sociales y
al Centro de Investigaciones Sociales, Económicas,
Políticas y Antropológicas (CISEPA),
por su apoyo en la gestión de este proyecto.

Contenido

Presentación

El presente material se ha elaborado gracias al financiamiento obtenido del fondo concursable 2011 del Vicerrectorado Administrativo de la Pontificia Universidad Católica del Perú. El objetivo de este material ha sido acercar, con un lenguaje sencillo, diversos conceptos estadísticos básicos a estudiantes de Ciencias Sociales y Humanidades dentro y fuera de la PUCP. Para esto, se ha tenido como estrategia diseñar un mapa mental web de la estadística básica, preparar videos para cada rama, y preparar un manual impreso y una wiki que los acompañen.

Cada parte del trabajo estuvo en manos de un alumno de la especialidad de Ciencia Política y Gobierno y los jefes de práctica que he tenido (quienes son parte del grupo de investigación del Laboratorio de Computación Social), lo que explica el estilo directo y sencillo de explicación que se aleja de la mayoría de materiales de estadística existentes. Este tour representa en sí la ruta que mis alumnos crearon para aprobar el curso de Estadística para el Análisis Político 1 (el que tiene cierta fama por su exigencia en nuestra especialidad). Se notará que hay ausencia de bibliografía y es que este trabajo lo han hecho utilizando sus notas de clase que aún guardan con nostalgia y que revivieron en este proyecto. Tuvimos además el agrado de contar con la revisión del doctor Jorge Aragón, profesor de nuestra especialidad en temas metodológicos. Sin embargo, cualquier error que aún exista es completamente atribuible a mi falta de verificación. En todo caso, esperamos sus recomendaciones a la dirección electrónica estadística.virtual@pucp.edu.pe puesto que, el objetivo de este material es que sea un material vivo, que aunque no pueda fácilmente modificarse en la versión impresa, si lo sea en la wiki y los demás componentes.

Espero que disfruten este aporte a su entrenamiento estadístico que redunde en su desempeño académico y laboral.

Prof. José Manuel Magallanes

Director del CISEPA

Introducción

La estadística es un área de estudio que cuenta con diversas técnicas y métodos que sirven de apoyo a distintas disciplinas interesadas en analizar la regularidad de sus objetos de estudio. Es de suma utilidad para la construcción de modelos que permiten verificar hipótesis y además, desde una perspectiva aplicada, apoyar en la toma de decisiones. El conocimiento matemático no es requisito obligatorio, pero sí el tiempo suficiente para practicar con los programas y ejemplos que hemos preparado.

En esta ocasión utilizaremos dos programas: Statistical Package for the Social Sciences (SPSS)[1] y R que serán nuestras principales herramientas para el análisis estadístico, las cuales se harán cargo de los cálculos matemáticos y de la construcción de los reportes numéricos y gráficos.[2]

Los contenidos se presentan en dos partes. La primera de ellas aborda lo que conocemos como análisis exploratorio. En esta sección aclararemos los conceptos de escala, codificación, recodificación, estadísticos y medidas de representación. La segunda unidad hace una breve introducción a la inferencia, tanto de forma univariada como bivariada.

Para el desarrollo de los distintos procedimientos estadísticos que abarca esta guía, se empleará una data especialmente preparada (ficticia). Es altamente recomendable que tengas tus propios datos para que trates de replicar lo que te mostremos. Asimismo, para que el aprendizaje resulte más ilustrativo se acompaña el contenido sobre estadística con la historia de Melissa Biondi, un personaje que nos ayuda a recrear el proceso de una investigación, desde la creación de una base de datos hasta las pruebas de hipótesis.

Es importante recordarte que este no es un material de autoaprendizaje, sino un material de apoyo a tus cursos de la Universidad.

---

1 También llamada PASW (*Predictive Analytics SoftWare*) desde 2009.

2 Los comandos y videos se pueden ver desde el mapa mental del material (ver pagina 67).

El presente material se ha elaborado gracias al financiamiento obtenido del fondo concursable 2011 del Vicerrectorado Administrativo de la Pontificia Universidad Católica del Perú. El objetivo de este material ha sido acercar, con un lenguaje sencillo, diversos conceptos estadísticos básicos a estudiantes de Ciencias Sociales y Humanidades dentro y fuera de la PUCP. Para esto, se ha tenido como estrategia diseñar un mapa mental web de la estadística básica, preparar videos para cada rama, y preparar un manual impreso y una wiki que los acompañen.

Cada parte del trabajo estuvo en manos de un alumno de la especialidad de Ciencia Política y Gobierno y los jefes de práctica que he tenido (quienes son parte del grupo de investigación del Laboratorio de Computación Social), lo que explica el estilo directo y sencillo de explicación que se aleja de la mayoría de materiales de estadística existentes. Este tour representa en sí la ruta que mis alumnos crearon para aprobar el curso de Estadística para el Análisis Político 1 (el que tiene cierta fama por su exigencia en nuestra especialidad). Se notará que hay ausencia de bibliografía y es que este trabajo lo han hecho utilizando sus notas de clase que aún guardan con nostalgia y que revivieron en este proyecto. Tuvimos además el agrado de contar con la revisión del doctor Jorge Aragón, profesor de nuestra especialidad en temas metodológicos. Sin embargo, cualquier error que aún exista es completamente atribuible a mi falta de verificación. En todo caso, esperamos sus recomendaciones a la dirección electrónica estadística.virtual@pucp.edu.pe puesto que, el objetivo de este material es que sea un material vivo, que aunque no pueda fácilmente modificarse en la versión impresa, si lo sea en la wiki y los demás componentes.

Espero que disfruten este aporte a su entrenamiento estadístico que redunde en su desempeño académico y laboral.

Prof. José Manuel Magallanes

Director del CISEPA

ESTADÍSTICA Y METODOLOGÍA DE LA INVESTIGACIÓN

El objetivo de este material es ayudarte a probar hipótesis simples con la estadística, tema que es de considerable relevancia dentro del proceso de la metodología de la investigación. Llegar a probar una hipótesis es sencillo con la ayuda de los programas estadísticos, pero existe una serie de pasos previos para lograr la formulación de una hipótesis adecuada. Esos pasos previos constituyen el esquema básico de cualquier investigación.

Melissa Biondi ha obtenido el puesto de asistente de investigación en la Dirección Psicopedagógica de Estudios Generales Letras. Este puesto recientemente creado tiene como función indagar sobre el rendimiento académico de los estudiantes de esta facultad. Cuando Melissa empezó su labor, se propuso trabajar con un tamaño grande de casos, de manera que le permitiera inferir a toda la población de Estudios Generales Letras.

Su proyecto fue aceptado y por tres meses tuvo la paciencia y el empeño de recopilar información sobre 99 estudiantes escogidos al azar que ingresaron el 2010. El lapso de tiempo que analizó fue de 2 años –4 semestres académicos– del 2010-1 al 2011-2. La base de datos que construyó sobre los 99 estudiantes tuvo información sobre los puntajes del examen de admisión y del examen de admitidos, las notas en todos los cursos, sus edades, la asistencia a clases, las escalas de pago, el nivel de inglés, si les gustaba la carrera o no y si al final de los dos semestres deseaban cambiarse a Estudios Generales Ciencias. Acompaña a Melissa en esta aventura y aprende junto con ella.

Generalmente, la investigación se promueve cuando se encuentran que las explicaciones aceptadas no son satisfactorias. La estadística descriptiva es importante en este momento pues nos informa la situación de diversos indicadores (sociales, psicológicos, políticos, educativos, etc.). Si el indicador de nuestro interés no se comporta como se esperaba, comenzamos a formular la pregunta de investigación; es decir, buscamos una explicación a lo que está sucediendo.

La formulación de una buena pregunta de investigación es primordial, dado que orienta el trabajo del investigador. En ella deberán estar incluidos los conceptos de interés que luego serán teóricamente sustentados en el marco teórico. Una vez delimitado el tema de investigación de manera teórica se podrá plantear la hipótesis, que es básicamente, la respuesta a la pregunta inicialmente formulada. En la hipótesis deben estar claramente expuestos los conceptos de interés y la relación existente entre ellos. A partir de allí solo queda contrastar tal hipótesis con lo que sucede en la realidad. Hecho el análisis respectivo se podrá reportar si la hipótesis era sostenible y se redactarán las conclusiones.

Dentro de la metodología de investigación debemos ser capaces de diferenciar algunos términos clave como: concepto, definición, variable, caso y valor. El concepto es un modelo mental (abstracción) que hace referencia a algo existente al otorgarle un nombre que permita identificarlo. La definición, por otro lado, es la explicitación del concepto que permite que este se diferencie de lo ya conocido. De ahí que la variable es simplemente una manera en que el concepto se manifiesta en el mundo y que puede tomar diversos valores o estados. El valor se obtiene mediante la medición o el conteo y el estado se obtiene mediante la observación y clasificación.

Por ejemplo, comparemos ‘democracia’ y ‘temperatura’. Ambos son ‘cosas’ diferentes que han ameritado que les demos nombre (concepto), pero el concepto ‘temperatura’ ya se puede definir unívocamente (conceptualización clausurada) como el promedio de energía cinética en la materia; sin embargo, el concepto ‘democracia’ no es un concepto clausurado, por lo que se siguen admitiendo diversas definiciones. La definición de temperatura permite que

se diseñe un instrumento confiable y válido para su medición (termómetro) y cuyos resultados ('nivel de temperatura') son datos que expresan su naturaleza de 'variable'[2] que se expresa a través de valores (en este caso numéricos). Las definiciones de democracia promueven diversas maneras de 'medirla', cada instrumento tiene una mayor o menor confiabilidad y validez. De ahí que esa medición presenta la variable 'nivel de democracia' a veces como un valor y otras como un estado. En ocasiones, el 'nivel de democracia' se conocerá mediante otras variables[3] ('nivel de justicia', 'nivel de libertad', etc.). Por ejemplo, la revista The Economist tiene una definición tal de democracia que usa las variables 'proceso electoral y pluralismo', 'libertades civiles', 'funcionamiento real del gobierno', 'participación política' y 'cultura política' que ayudan a construir la variable 'nivel de democracia'. El valor o estado de la variable siempre corresponde a alguna unidad de estudio[4]: la temperatura es una variable de la unidad de estudio 'persona', por lo que habrá un valor por persona. Cuando tengas un tamaño considerable de unidades de estudio y valores por variable comenzaremos el análisis estadístico.

Melissa Biondi no tenía una pregunta de investigación a la cual responder con una hipótesis, tenía varias: ¿existe diferencia según la escala de pago en los cursos de matemática? ¿Los que tienen buenas calificaciones en los cursos de filosofía también tienen buen puntaje en los cursos de historia? ¿Los que tienen bajo nivel de inglés son los que obtuvieron bajo puntaje en el curso de Quechua? ¿Las mujeres asisten más a clases que los hombres? De manera que su proyecto era más ambicioso, quería encontrar diferencias y correlaciones entre las variables que ella planteaba con el fin de brindar asesoría personalizada o talleres de estudio a los estudiantes que lo requerían, y de la misma forma, proponer algún mecanismo de incentivos para que los que tenían buenas calificaciones se sigan esforzando.

1 También llamada PASW (*Predictive Analytics SoftWare*) desde 2009.

2 Para diferenciarlo de 'constante'.

3 Cuando este sea el caso, estaremos frente a una variable latente, es decir, aquella que se mide mediante otras variables observadas (no latentes). De ahí que, según la definición por la que optemos habrá que buscar o recolectar varios valores.

ANÁLISIS EXPLORATORIO: ¿CÓMO ESTÁN LOS DATOS CON LOS QUE VAMOS A TRABAJAR?

Explorar significa conocer el comportamiento de las variables de las unidades de estudio. Para conocer este comportamiento se tienen pasos sencillos, pero cada uno amerita mucha paciencia y reflexión, los cuales desarrollamos a continuación.

1. PASO 1: IDENTIFICACIÓN DE LA ESCALA DE LA VARIABLE

Las técnicas estadísticas se aplican según la escala en que se encuentren los datos y estos datos son o valores o estados de la variable para cada unidad de estudio. La correcta identificación de la escala en la que se presentan los datos es clave para todo lo demás.

1.1. ESCALAS PARA VARIABLES CUALITATIVAS

Estas variables presentan datos que informan de los estados de las unidades de estudio. Estos estados se organizan en dos escalas: escala nominal y escala ordinal.

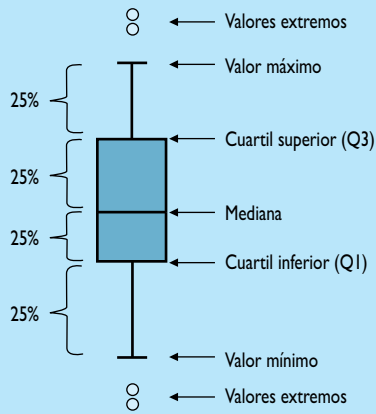
En la escala nominal, sus estados (también llamados modalidades o categorías) no presentan orden entre sí. Un ejemplo puede ser la variable 'signo del zodiaco' (que tendrá respuestas como 'Sagitario', 'Aries', etc.), o 'medio que utiliza para informarse' (con respuestas como 'Internet', 'radio', 'televisión', etc.). Como vemos, los estados de estas variables no tienen orden entre sí. De acuerdo al número de estados, la escala nominal puede ser: dicotómica (dos categorías u opciones) o politómica (más de dos categorías). Los gráficos básicos que se pueden realizar para las variables nominales son

[5] Gráficos explicados en la página 20

los diagramas de barras y los gráficos de sectores.[5] Estos permiten visualizar de manera clara las frecuencias de las categorías/modalidades de las variables nominales.

Los datos en escala ordinal presentan respuestas que representan estados ordenados (en niveles), pero entre niveles consecutivos no hay distancia numérica. Ejemplo común es la variable 'nivel educativo' (puede contener las categorías 'primaria completa', 'secundaria completa' y 'superior completa'). Para este tipo de datos también se pueden utilizar los diagramas de barras y los gráficos de sectores. Pero existe otro gráfico que resume mejor una variable ordinal llamado diagrama de cajas o boxplot (Gráfico 1). Esta ayuda visual presenta información sobre las medidas de posición (cuartiles); es decir, divide los valores en cuatro partes donde cada parte contiene el 25% de casos. Al mismo tiempo, este gráfico, nos da información sobre la tendencia central, dispersión y simetría de los datos de estudio. Además, este gráfico permite identificar con claridad casos que se alejan de manera poco usual del resto de los datos, a estas observaciones se les conoce como valores extremos y atípicos.

Gráfico 1. Boxplot o Diagrama de Cajas.



Elaboración propia.

En la base de datos de Melissa, llamada 'Letras.sav' se pueden ver todas las variables que ella ha podido consignar, sus respectivas etiquetas, así como las respuestas codificadas de sus 99 casos. Ella tiene 15 variables categóricas, 6 nominales y 9 ordinales. La variable 'código del alumno' es de tipo cadena (porque no es número) y permite la identificación de los casos (ver Tabla 1).

Tabla 1. Variables en archivo Letras.sav

Variables categóricas nominales	Variables categóricas ordinales
• Sexo • Salón de letras • ¿Le gustó la universidad en su primer año? • ¿Se cambiaría a ciencias en este primer año? • ¿Le gustó la universidad en su segundo año? • ¿Se cambiaría a ciencias en este segundo año?	• Asistencia al semestre 2010-1 • Escala económica en el primer semestre • Asistencia al semestre 2010-2 • Escala económica en el segundo semestre • Asistencia al semestre 2011-1 • Escala económica en el tercer semestre • Asistencia al semestre 2011-2 • Escala económica en el cuarto semestre • 'Nivel de inglés'

Elaboración propia.

1.2. ESCALAS PARA VARIABLES CUANTITATIVAS

En los casos anteriores, hemos hablado de variables que se representan mediante estados y ahora abordaremos el estudio de las variables cuantitativas. Cuando hablemos de este tipo de variables nos referimos a que sus datos están en escala numérica;[6] por lo que estamos haciendo referencia a la idea de magnitud. Esta variable, como las anteriores, tiene también subdivisiones. En este caso pueden ser discretas (conteos) como número de hijos, o continuas (medición), que admite decimales como, 'peso', 'altura', etc.

[4] También Unidad de Análisis, Unidad de Información, etc.
[5] Gráficos explicados en la página 11.
[6] Se suele diferenciar entre escala de intervalo y escala proporcional, pero esa diferencia no la utilizaremos en este manual.

Podemos hacer distinciones que nos permitan identificar las distintas variables; sin embargo, la escala de las variables no está definida a priori. Esto quiere decir que la escala no es intrínseca a una variable cuantitativa o cualitativa; sino que, esta es determinada por el propio investigador. Por ejemplo, el concepto educación, puede ser medido en las tres escalas que hemos mencionado. Para el caso de la escala nominal, el concepto 'educación'

2. PASO 2: DETALLES A CONSIDERAR CON LAS ESCALAS

Luego de identificar las escalas podremos ver qué tenemos realmente en nuestros datos. Ahí estaremos conscientes que hay una serie de procedimientos pendientes antes de hacer un análisis estadístico. Veamos estos temas a continuación.

Melissa ha consignado como variables numéricas a las notas de 20 cursos, al examen de admisión y al examen de admitidos. Estas variables son numéricas discretas y pueden optar cualquier valor de 0 al 20. Las variables 'examen de admisión' y 'examen de admitidos' pueden tomar cualquier valor de 0 a 1000 y también son discretas.

2.1. CODIFICACIÓN

La codificación es un paso importante para que los programas informáticos traten los datos. Las computadoras son más eficientes en el tratamiento estadístico si los datos que manejan son números. Por ello, cuando se abren y revisan algunas bases de datos en una computadora, todo lo que se observa son números, aun cuando solo algunos de esos números representan variables en escala numérica y las demás representan variables en escala nominal u ordinal. Así, en vez de señalar literalmente el nivel educativo, aparecen números que indican algún nivel de educación (1 para 'primaria', 2 para 'secundaria', etc.).

Solo en el caso de las variables en escala numérica, esos números representan efectivamente una magnitud. Así, si la variable 'medio de información preferido' tiene el valor 2 para 'radio' y el 4 para 'televisión' no implica que televisión sea el doble o más importante que radio, aquí esos números son solo una simple etiqueta que ayuda a diferenciar las dos categorías. Es diferente en el caso de la variable 'número de hijos', donde la persona que tiene 4 hijos efectivamente posee más hijos que aquella que tiene 2 y de hecho tiene el doble por ser una magnitud real.

Otro uso particular e importante de los códigos son los valores perdidos (*missing values*), que son los códigos que se utilizan para indicar respuestas inadecuadas, inapropiadas o faltantes, pero que se indican de manera explícita. Estos valores no se utilizan en los cálculos, de hecho la codificación que tienen permite que los programas informáticos los ignoren. Normalmente, por convención a los valores perdidos se les otorga el valor '99', '999' o se deja la celda vacía.

Melissa ha codificado también las variables categóricas de su data, tanto las ordinales como las nominales. Por ejemplo, la variable 'escala económica en el primer semestre' es ordinal y va del 1 al 5 donde 1 representa la 'escala económica más baja', 2 la 'escala baja', 3 la 'escala intermedia', 4 'la escala alta' y 5 'la escala económica más alta'. Para el caso de una nominal, la variable '¿le gustó la universidad en su primer año?' tiene como etiqueta 'no le gustó' al valor 2 y 'sí le gustó' al valor 1.

2.2. TRANSFORMACIÓN

Muchas veces es necesario realizar modificaciones a algunas variables de la data con la que estamos trabajando. Existen dos casos en los que se emplea la re-codificación los cuales serán explicados a continuación.

2.2.1. Organizar intervalos

Tal como explicamos en el apartado anterior, las variables numéricas –sean conteos o mediciones– representan magnitudes reales, y por esto, al hacer el recojo de información el número de posibles respuestas puede ser bastante alto. Esto trae algunos problemas cuando el investigador desea mostrar una tabla de frecuencias, puesto que, existirían tantas filas como respuestas distintas. Por esto, es sumamente útil recodificar la variable en intervalos para facilitar la lectura de los datos.

Por ejemplo, si en una data con 1000 casos contamos con la variable 'ingresos', es muy probable que existan tantas respuestas distintas como casos. Entonces, una alternativa es recodificar la variable en intervalos. Por lo general, al valor máximo de la variable se le resta el valor mínimo y se divide entre el número de intervalos que el investigador crea conveniente. De esta forma se obtiene la amplitud intervalar, la cual permite crear intervalos regulares. En el caso de la variable ingresos, si nuestro mayor valor es 5000 y el menor valor es 1000, tomando en cuenta que deseamos crear 5 intervalos, la amplitud de cada intervalo será 800, por lo que los intervalos quedarían de la siguiente manera:

[1000; 1800]

]1800; 2600]

]2600; 3400]

]3400; 4200]

]4200; 5000]

Tras ello, obtenemos cinco intervalos y si a cada grupo le asignásemos alguna categoría habríamos transformado la variable numérica a escala ordinal.

2.2.2. Cambio de monotonía

En muchos casos, las variables ordinales que presenta nuestra data no poseen la misma monotonía; es decir, algunas se encuentran codificadas de forma ascendente y otras tantas de forma descendente. Esto podría parecer un simple detalle; sin embargo, es conveniente trabajar con una sola monotonía para evitar problemas en los cálculos de pruebas estadísticas o para la lectura de los resultados. Es además sumamente importante para el desarrollo de

índices. Por ejemplo, la variable nivel educativo podría estar codificada de la siguiente manera: Superior '1', Secundaria '2' y Primaria '3'. En cambio, la variable nivel de inglés podría estar codificada de esta otra forma: básico '1', intermedio '2' y avanzado '3'. Tal como podemos ver, la primera, nivel educativo, está codificada de forma descendente, mientras que la segunda variable, nivel de inglés se encuentra codificada de forma ascendente. Para evitar inconvenientes el investigador debería recodificar la primera variable -nivel educativo- y cambiar su monotonía. De esta forma la variable recodificada quedaría así: Primaria '1', Secundaria '2' y Superior '3'.

En la data que ha preparado Melissa se tiene variables ordinales que provienen de variables numéricas, estas hacen referencia al porcentaje de asistencias a lo semestres, y hay cuatro de este tipo. Ella ha calculado la asistencia promedio de los cinco cursos que cada estudiante ha llevado en cada semestre y de acuerdo al porcentaje de asistencias que ha tenido los ha clasificado en cuatro grupos. No necesariamente los ha clasificado por intervalos iguales, sino según un criterio relevante para sus intereses analíticos. De manera que el valor 0 lleva la etiqueta 'hasta el 80% de inasistencias', el 1 'hasta el 50% de inasistencias', el 2 'hasta el 20% de inasistencias' y el 3 'hasta el 3% de inasistencias'. Como en su data no ha habido alguien con más del 80% de inasistencias no ha tenido que crear un intervalo adicional.

3. PASO 3: CÁLCULO DEL VALOR REPRESENTATIVO Y LA EVALUACIÓN DE SU CALIDAD

El valor representativo es clave en el análisis estadístico,[7] pues justamente se usan las medidas estadísticas (también llamadas estadísticos o estadígrafos) para resumir en algún valor el comportamiento promedio de las unidades de estudio. La importancia de haber identificado la escala permitirá decidir qué estadístico se le debe calcular a la variable de interés y debe haberse entendido la codificación encontrada y realizar ajustes en la monotonía de la variable de ser necesario. Si esto no se verificado y trabajado antes, no se deben hacer cálculos aún.

[7] Se le conoce a menudo como el 'valor central' o 'medida de centralización'.

Luego de haber sistematizado su data, Melissa se decide a obtener las primeras medidas y gráficos de sus variables. Considera esta parte importante porque le permite entender mejor el comportamiento de los estudiantes de Letras (por ejemplo si hay más tardones que puntuales, si hay más estudiantes con un nivel básico del inglés que con uno alto, etc).

3.1. PARA DATOS EN ESCALA NOMINAL

Para este caso el estadístico es la moda. La moda informa cuál es el estado de la unidad de análisis que más se repite. Su identificación es muy sencilla si se ha organizado a las variables en una tabla de frecuencias, como se ve en la Tabla 2. Ahí, para la variable sexo, se nota que la moda es 'mujer', pues es el estado que se repite más veces (53 versus 46).

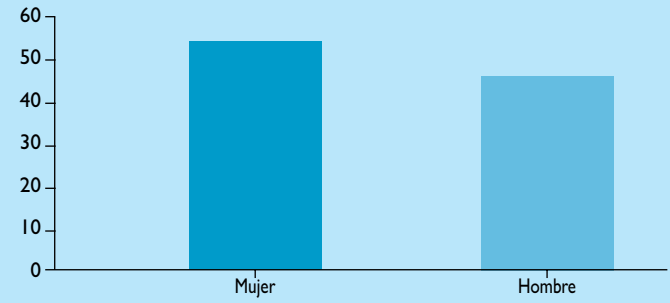
Para ver la calidad representativa de la moda se debe identificar cuánto representa porcentualmente, y según eso decidimos su calidad representativa. En la Tabla 2 se ve que la moda es mujer pero que es 53%, ¿eso hace que la moda sea representativa? ¿Sería mejor si la moda fuera el 90%? Es muy importante recordar que cuando se informa el valor representativo, en este caso la moda, los receptores de la información (maestros, políticos, padres de familia, médicos, entre otros) necesitan saber más detalles de esa 'representatividad'; por lo tanto, en cada situación, según la necesidad, se podrá interpretar la moda valorando a la vez este valor porcentual. Para el caso de los resultados electorales basta

Tabla 2. Tabla de frecuencias de la variable 'sexo'.

		Sexo			
		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Mujer	53	53.5	53.5	53.5
	Hombre	46	46.5	46.5	100.0
	Total	99	100.0	100.0	

Elaboración propia.

Gráfico 2. Gráfico de barras de la variable 'sexo'.

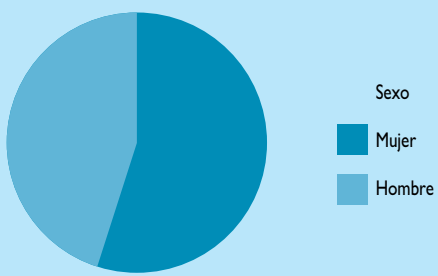


Elaboración propia.

que la moda sea más que el 50%, pero si alguien obtiene 49% y los demás competidores obtienen 21% y 30%, la normatividad peruana indica que esa moda, aun cuando representa quién obtuvo más votos, no cumple con la calidad representativa necesaria para ser declarado Presidente.

Esto también se puede apreciar gráficamente con la representación de barras (Gráfico 2) y el gráfico de sectores (Gráfico 3). Como veremos ambos presentan el aporte de cada categoría, al total lo que ayuda a comparar los valores entre categorías.

Gráfico 3. Gráfico de sectores de la variable 'sexo'.



Elaboración propia.

Con estos dos gráficos y la tabla de frecuencias, Melissa ya tiene para abrir la primera parte de su informe sobre la composición de su muestra estudiantil de Estudios Generales Letras. Habría más mujeres que varones, pero no puede inferir todavía porque no sabe si la diferencia existente entre ambos grupos es significativa; es decir, afirmar con un cierto nivel de confianza que esa diferencia se mantiene en la población.

3.2. PARA DATOS EN ESCALA ORDINAL

El segundo caso en variables categóricas es el de las variables en escala ordinal, en cuyo caso utilizamos el estadístico conocido como mediana[8] para informar cuál es el valor representativo. Si se hiciera a mano este cálculo, esta medida requeriría que se ordenen los estados de menor a mayor para obtener un punto de corte (centro) que depende de la cantidad de elementos para su cálculo. Por ejemplo, si se tiene 99 elementos, el valor del elemento (o caso) 50 será la mediana (se elige este caso ya que hay 49 elementos a la izquierda y a la derecha). Entonces, mientras los elementos estén ordenados según su valor, la mediana siempre será el valor del elemento central, sin importar que tan alejados estén los demás valores de este.

Para ver la calidad representativa de la mediana se calcular el IQR.[9] Veamos un ejemplo:

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
• Grupo 1	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	3	3	3	3	4	4	4	5	5	5	5	6	6	6
• Grupo 2	1	1	1	1	1	1	2	3	3	3	3	3	3	3	3	4	4	4	4	4	4	4	5	5	6	6	6	6	6
• Grupo 3	1	2	2	2	3	3	3	4	4	4	4	4	4	4	5	5	5	5	5	6	6	6	6	6	6	6	6	6	6
								q1							md							q3							

[8] La moda también es calculable aquí, pero se prefiere a la mediana para aprovechar las características de las variables ordinales.

[9] *Inter Quartil Range* o Rango intercuartil, que es el cuartil 3 menos el cuartil 1.

Como se aprecia, hay tres grupos de datos con diferentes estados[10]. Cada grupo va del 1 al 6 (esto podrían ser tres series de respuestas en una escala Likert), pero cada grupo tiene 'personas' (unidades de estudio) que en conjunto presentan patrones de respuestas diferentes. Ahí señalamos las medianas con md y como cuartiles[11] a q1 y q3 (la mediana tiene la propiedad de ser también el cuartil 2). Así, los IQR para cada grupo son 2,1 y 2. Por lo que la mediana del grupo 2 sería la mejor.

En todo caso, hay que tener siempre en cuenta que la mediana siempre indica el valor o estado debajo del cual está el 50% de la población en la variable de análisis. Eso convertiría a la mediana en un estadístico muy robusto, inclusive para los datos de la escala numérica, salvo que la media, la cual veremos a continuación, tenga mejores medidas de calidad.

3.3. PARA DATOS EN ESCALA NUMÉRICA

En las variables numéricas se pueden calcular la moda y la mediana, pero el valor representativo más apropiado para esta escala es la media.[12] Esta es calculada sumando los valores de cada unidad de estudio y dividiendo el resultado obtenido por la cantidad de unidades de estudio. La media es mejor que la mediana en el sentido que utiliza todos los valores para su cálculo y la mediana utiliza las posiciones para su cálculo. Pero la mediana siempre nos dice el valor a la mitad de la distribución y la media lo hará si es de calidad. Enfaticemos estas diferencias entre media y mediana, para ello imaginemos que tenemos cinco grupos (A, B, C, D y E) con 11 casos[13] cada uno como se muestra en la Tabla 3.

---

[10] Recuerde que estamos en escala ordinal, por lo que no decimos valores.

[11] El cuartil uno muestra el valor o estado que a lo más llegaría el 25% más bajo de la población. El cuartil tres muestra el valor o estado que a lo más llegaría el 75% más bajo de la población, o el mínimo valor o estado del 25% más alto de la población. Los cuartiles son parte de las medidas de posición.

[12] En este material se hablará solo de la media aritmética, no teniendo en cuenta la media geométrica ni la media armónica.

[13] Unidades de estudio.

Tabla 3. Comparación media-mediana.

Caso:	1	2	3	4	5	6	7	8	9	10	11	Media	Mediana
• Grupo A	30	30	30	30	30	30	30	30	30	30	31	30.09	30.00
• Grupo B	30	30	30	30	30	30	30	30	30	30	50	31.82	30.00
• Grupo C	29	30	30	30	30	30	30	30	30	30	30	29.91	30.00
• Grupo D	20	20	20	20	20	20	30	30	30	30	30	25.45	20.00
• Grupo E	20	20	20	20	20	24	30	30	30	30	30	24.91	24.00

Elaboración propia.

En la Tabla 3 podemos comprobar que la media se calcula sumando todos los valores y dividiendo por la cantidad de elementos (en este caso siempre se divide entre 11). En las primeras 3 filas (ejemplos A, B y C) cambia el valor mayor o el menor y se observa que la media se altera. Sin embargo, la mediana sigue siendo la misma. La mediana no se altera por los valores, sino que depende de la posición de los elementos para su cálculo. En este caso, como hay 11 elementos, el elemento 6 está al medio (hay la misma cantidad de elementos a la izquierda que a la derecha). Entonces, mientras los elementos estén ordenados según su valor, la mediana siempre será el valor del elemento central, sin importar que tan alejados estén los demás valores de este.[14] De ahí que, en nuestro ejemplo, la mediana simplemente está tomando el valor que encuentra en el valor del elemento 6.

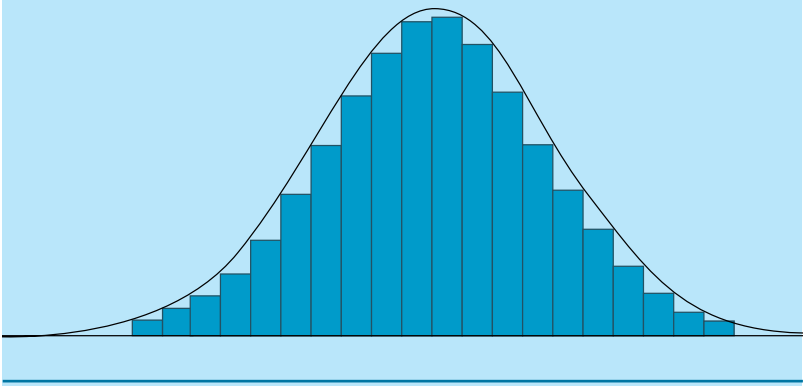
Como vemos, ambas medidas quieren informar lo mismo (el valor representativo o medida central de esa variable), pero cada una utiliza un proceso diferente para ello. Si la media y mediana coinciden, quiere decir

[14] El ejemplo muestra un número impar para hallar la mediana, pero si el número fuera par se sacaría el promedio de los dos valores que se encuentran en el centro o se aplicaría alguna otra técnica alternativa para encontrarlo.

que podemos confiar que el valor de ambos representa bien a ese grupo. Pero si se alejan cada vez más quiere decir que hay un sesgo; es decir, el valor representativo obtenido no es muy informativo. En otras palabras, en caso de alejarse no hay un valor representativo claro para esa variable, por lo que apostar por la mediana es más recomendable.

El histograma (Gráfico 4) es una representación gráfica de la distribución de los valores en una variable cuantitativa que se muestra en forma de barras seguidas. Este gráfico nos permite describir el comportamiento de un conjunto de datos; es decir, el histograma nos permite identificar cuantas veces se repite un mismo valor, así como la frecuencia con la que se presenta.

Gráfico 4. Histograma.



Elaboración propia.

Para desarrollar con mayor profundidad el análisis de las variables numéricas los programas cuentan con otros estadísticos que nos permitirán hacer un mejor análisis. Estos otros estadísticos se muestran en la Tabla 4, y nos servirán justamente para explicar la calidad de la media.

Tabla 4. Otros estadísticos para datos en escala numérica.

• Valor representativo		Media	19.02
• Valor representativo alternativo		Mediana	19.00
• Dispersión		Desviación típica	1.726
		Varianza	2.979
		Coeficiente de variación	0.095
		Mínimo	16
		Máximo	22
• Forma	Asimetría	Asimetría	-0.105
		Error típico de asimetría	0.243
	Curtosis	Curtosis	-0.826
		Error típico de curtosis	0.481

Elaboración propia.

Tener un mejor veredicto de la calidad de la media requiere mayor análisis de la dispersión y la forma de los datos, lo que enriquece y ala vez complejiza el análisis para el caso particular de la escala numérica. Veamos cada uno con detalle a continuación.

3.3.1. Dispersión

La dispersión nos da una señal de que tan representativo puede ser un valor central. Las principales medidas de dispersión son la varianza (σ^2) y el coeficiente de variación. Además, de la varianza podemos obtener la desviación típica (σ) o la desviación estándar, que es su raíz cuadrada.

La idea de la varianza (o desviación típica) es muy simple: su valor desea mostrar que tan alejados están los datos de la media. Sin embargo, la varianza no tiene valores topes por lo que su uso es más importante para comparar la misma variable en diferentes grupos. Por ejemplo, en el caso de los valores de la Tabla 4, si los datos representasen las edades del grupo A de 'cachimbos' de educación, si el grupo B tuviera la misma media, pero su varianza fuese 1.5, podríamos afirmar que la media del grupo B es de mejor calidad informativa. Pero si no hubiera grupos para comparar los estadígrafos

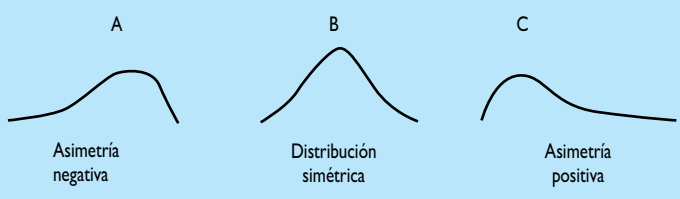
de dispersión de la Tabla 3, no podríamos afirmar con certeza el nivel de dispersión. Lo mismo ocurre con la desviación típica, pero esta medida tiene la ventaja de estar es la misma unidad de medida que la variable.[15]

El coeficiente de variación (CV) mejora a la varianza y a la desviación típica en que no solo permite comparaciones de la dispersión entre la misma variable numérica para diferentes grupos, sino que permite comparar la dispersión entre cualquier par de variables numéricas. El CV es el cociente entre la desviación estándar y la media de la distribución, lo cual da por resultado un cociente que no tiene unidades y permite la comparación entre diferentes distribuciones. A mayor dispersión (mayor desviación típica) el coeficiente saldrá mayor.[16] El CV tiene la ventaja de oscilar entre 0 y 1, pero para su mejor lectura debe convertirse a porcentaje multiplicándose por 100%, quedando en el analista definir el máximo CV a aceptar (generalmente, 0.1 o 10%).

3.3.2. Asimetría

La asimetría nos brinda información de la distribución de los valores de la variable, pero solo concerniente si este tiene algún sesgo, como se muestra en el Gráfico 5.

Gráfico 5. Tipos de simetría.



Elaboración propia.

[15] Es decir, si la variable es edad, la varianza indica la dispersión en edades al cuadrado.

[16] La media no debe ser cero pues hace no calculable al CV. EL CV es recomendable cuando la variables o lo puede tomar valores positivos (para evitar cambio de rango).

Las tres curvas que mostramos ejemplifican las posibles distribuciones que se pueden generar.

La primera curva (A) indica que los valores mayores de la variable son los más comunes o abundantes. A esta forma, como se ve en el gráfico se le denomina asimetría negativa o a la izquierda. Puede suceder que en una muestra recogida en una universidad la mayoría tenga más de 15 años de educación y que una minoría tenga pocos años estudiando. Consideraremos que una variable tiene asimetría negativa cuando el valor obtenido en el estadígrafo de asimetría lleve un signo negativo. Como dato adicional, ante esta situación asimétrica, se cumple que la media es menor que la mediana, y esta es menor que la moda.

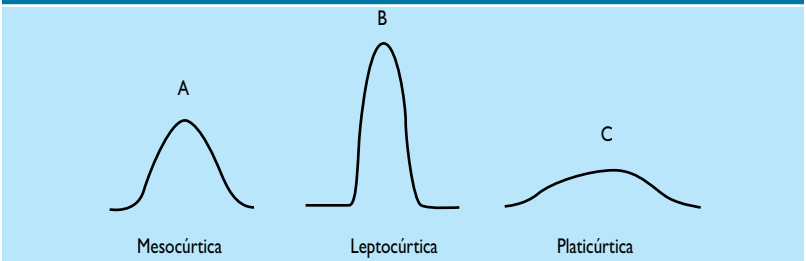
La segunda curva (B) nos muestra una mayor frecuencia de los valores medios y menor a los extremos. En las distribuciones simétricas, la moda, mediana y media coinciden. Es un poco difícil encontrar una distribución perfectamente simétrica en la vida real aunque podemos hipotetizar que si preguntamos las edades en un salón de quinto año de secundaria, las edades oscilarán entre 15 a 17 años en mayor medida, siendo otras edades de menor ocurrencia. Para reconocer una variable con distribución simétrica la medida de simetría tiene que tender a cero. Tener una distribución normal nos acerca a contar con una distribución especial conocida como la normal, la que es requisito para varias pruebas estadísticas que veremos más adelante.

La tercera curva (C) nos indica que los valores más bajos de la variable son los más abundantes. Esta curva representa lo que se denomina técnicamente una asimetría positiva (los datos están sesgados a la derecha). ¿Qué tipo de variable tendrá esta forma? Quizá una muestra recogida de los sueldos de la gente en un distrito de la sierra de Ayacucho bajo la línea de pobreza, pues se espera que en general haya mucha gente que gane poco y que los que ganan más que el común de la gente sean muy pocas personas. Cuando una variable tiene una distribución parecida a esta figura, la asimetría sale mayor que cero. Además, en la asimetría positiva se cumple que la media es mayor que la mediana, y esta es mayor que la moda.

3.3.3. Curtosis

Este es otro estadístico que informa sobre la forma de la distribución de la variable y nos indica qué tan 'aplanada' (o empinada) está la curva que representa esa distribución (Gráfico 6). Si el valor saliera positivo alejándose mucho del cero estaríamos seguros de que se trata de una curva empinada. Si el valor saliera negativo alejándose mucho del cero esperaremos una bastante plana.[17] Cuando una curva es muy empinada se le dice leptocúrtica, cuando está bien aplanada se le denomina platicúrtica, y cuando adquiere una forma acampanada (cuando el valor del índice de curtosis es o se acerca a 0) nos referimos a una curva mesocúrtica.

Gráfico 6. Tipos de curtosis.

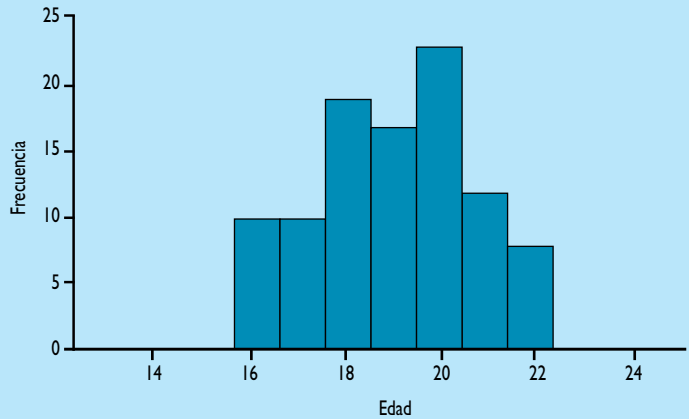


Elaboración propia

Los estadísticos de asimetría y curtosis pueden ser apreciados a través del histograma que, como ya habíamos revisado, es el principal gráfico para entender el comportamiento de una variable numérica (Gráfico 7). Es parecido al gráfico de barras, pero es aplicable solo a las numéricas puesto que visualiza la distribución.

[17] La palabra 'mucho' no es concreta y el investigador verá de utilizar pruebas auxiliares para verificar la Curtosis. En el SPSS, la Curtosis aparece además con su error típico, y se suele tener como regla básica que sí habrá desvío considerable si el estadístico de Curtosis dividido entre su error típico es mayor que 2 en valor absoluto (sin importar el signo). Lo mismo se aplica los valores del coeficiente de asimetría.

Gráfico 7. Histograma de la variable numérica edad.



Elaboración propia.

Para analizar la información de los estadísticos y la gráfica podemos proceder presentando la media que es 18.9, indicando los valores mínimo y máximo (16 y 22), hasta aquí podemos pensar que la media es muy buena, entonces vemos la gráfica y nos damos cuenta de que no parece una campana, y más bien parece que la asimetría es negativa: la media es menor que la mediana (sería al revés si fuera positiva).[18] Para ver si eso es excesivo dividimos la medida de asimetría entre su error típico obteniendo -0.105 ÷ 0.243 = -0.43. Como el valor absoluto es menor que 2, concluimos que esa diferencia entre media y mediana no es tan significativa, por lo que la media sigue siendo objeto de interés (sino solo interesaría la mediana). Aunque la asimetría no sea problema, podría serlo la dispersión o alejamiento de los demás valores

[18] La media es muy buena cuando el histograma muestra que la distribución de los valores es simétrica, mesocúrtica y unimodal (una sola moda).

de la media, por lo que viendo la desviación estándar de 1.726 nos animamos a calcular el coeficiente de variación, que nos da 1.726 ÷ 19.02 = 0.095, que convertido en porcentaje da: 9%; es decir, la variabilidad de los datos es del 9%. Si el investigador considera que esta es una variabilidad aceptable, la media aun no pierde poder informativo. Antes de decidir ello, revisamos la curtosis, pues si esta es significativamente negativa comprometería al coeficiente variación. Así, calculamos -0.826 ÷ 0.481 = -1.71 viendo que no es significativamente platicúrtica. Tenemos así que la dispersión tampoco es problema, pero aun así miramos de nuevo la gráfica, identificando que los valores menores tienen pocos casos (hay varias barras menor que 19 y no decrecen hacia la izquierda y la altura de sus barras no es baja), poniendo otra vez en duda la representatividad de la media (después de todo, queremos decir la edad que represente al grupo y quizá una edad más joven sea más pertinente). Sin embargo, hay que tomar una decisión, y por lo que hemos visto, hay mucha evidencia a favor de la media, pues aunque hay muchos casos menores que 19, hay un pico de alumnos con 20 también, por lo que mantener como edad representativa al 19 parece razonable.

Con los estadísticos y los gráficos obtenidos Melissa ya puede saber la composición etaria de su muestra. De los 99 estudiantes se tiene que en promedio tienen 19.02 años y que sus edades oscilan entre los 16 y los 22 años. Con las medidas de asimetría y curtosis refuerza la idea de que el valor representativo de su variable es la media ya que la curva de la distribución no está tan sesgada ni a la izquierda ni a la derecha, y a su vez, tampoco es tan alargada ni tan aplanada. Con una distribución centrada y mesocúrtica Melissa tiene seguridad que la moda, la media y la mediana no están tan alejadas y que la distribución de la variable edad se acerca a una curva normal.

Para su informe Melissa ha explorado todas las variables y les ha sacado sus medidas correspondientes. Antes de entrar a la parte de ejercicios anímate a hacer lo mismo y guíate con los videos.

Ejercicios 1

1. **¿Cuál de los cursos que se llevaron el primer ciclo tiene más puntaje en promedio?**
2. **Marque si es correcto:**
   a. La media del curso de matemática 1 es 14.
   b. La nota que más se repite es 16.
   c. La mediana es la misma en los 3 cursos de historia de letras.
   d. Existe mucha variación en los cursos de matemática.
   e. Hay más alumnos en las dos primeras escalas del primer semestre.
   f. Hay más desaprobados en los dos cursos de matemática que en el resto de cursos.
   g. El curso de apreciación musical es el que agrupa las mejores calificaciones en el semestre.
   h. El curso de lógica es el que tiene más desviación típica tiene.
3. **¿Qué cursos se distribuyen de manera normal en el primer ciclo?**
4. **¿Cuántos cursos del segundo semestre son los que tienen una distribución sesgada?**
5. **Se suele decir que:**
   a. Los cursos de matemática tienen cada uno un 30% de alumnos desaprobados.
   b. El curso de Apreciación Musical es el que tiene más gente aprobada.
   c. Hay más desaprobados en el primer ciclo comparado con el segundo
6. **El coeficiente de variación y rango son medidas de:**
   a. Dispersión
   b. Centralización

c. Posición

d. Concentración

7. Con el boxplot podemos ver la …… de la variable ……

a. Curtosis y simetría – nominal

b. Posición, simetría, dispersión – ordinal y numérica

c. Concentración, posición y apuntamiento – ordinal y numérica.

d. Dispersión y concentración – ordinal, numérica y nominal

8. La barra de error con la desviación estándar es una gráfica que:

a. Visualiza la distribución de la variable

b. Visualiza la dispersión de la media de la variable

c. Visualiza la concentración y dispersión de la variable

d. Visualiza la simetría y dispersión de la variable.

9. El histograma permite ver, marque más de una:

a. El apuntamiento de la variable ordinal.

b. La concentración y dispersión de la variable numérica discreta.

c. La simetría y apuntamiento de la variable numérica continua.

d. La concentración de la variable numérica discreta.

10. El boxplot no muestra, marque más de una:

a. El rango intercuartílico de la variable ordinal.

b. La simetría de la variable numérica.

c. La mediana como medida de centralización.

d. El sesgo de la media

11. De las medidas de centralización, marque más de una:

a. La media es la medida más sensible a los valores extremos.

b. La mediana es la medida menos sensible a los valores atípicos.

c. La moda es una medida tanto para variables numéricas y categóricas

d. La media pierde representatividad si la distribución es sesgada.

ANÁLISIS INFERENCIAL: ¿QUÉ INFORMACIÓN PODEMOS OBTENER DE LOS DATOS?

Aquí nos toca introducir temas de estadística inferencial, lo cual nos permitirá deducir propiedades de una población a partir de una muestra. Es importante recalcar que los resultados obtenidos a través de esta técnica son probabilísticos por lo que dependen de la representatividad de la información obtenida.[19] En lo que queda del material, trabajaremos la inferencia univariada y bivariada.

4. INFERENCIA UNIVARIADA

Los estadísticos de inferencia univariada permiten saber si lo hallado en una variable de una muestra es suficiente para conocer mejor el comportamiento de esa variable en la población de interés.

Básicamente en la inferencia univariada queremos saber dos cosas:

1. Si los valores de nuestra muestra se distribuyen de manera similar a algún modelo de distribución estadístico teórico.
2. Si los estadísticos o estadígrafos de representatividad, revisados en el primer capítulo, pueden ser una buena estimación del parámetro de la población.

La inferencia univariada varía de acuerdo a la escala de medición. Por ello, presentaremos las pruebas inferenciales según el tipo de variable.

[19] En este material suponemos que tienes tus datos, y que estos han sido obtenidos de un adecuado proceso de muestreo.

4.1. PARA DATOS CATEGÓRICAS

Para este tipo de variables queremos saber si es posible que las proporciones de las categorías presentadas en nuestra muestra se mantengan en nuestra población.

4.1.1. Prueba binomial

Usaremos esta prueba en el caso de tener variables categóricas dicotómicas (con solo dos categorías posibles). Por ejemplo, en la Tabla 5 queremos saber si es probable que en la universidad exista la misma proporción para ambos sexos (50% mujeres y 50% hombres).

Tabla 5. Prueba binomial de la variable 'sexo'.

		Categoría	N	Proporción observada	Proporción de prueba	Significancia
Sexo	Grupo 1	Mujer	53	0.54	0.50	,547[a]
	Grupo 2	Hombre	46	0.46		
	Total		99	1.00		

Elaboración propia.

Para facilitar la lectura de los resultados explicaremos a continuación la convención a seguir que nos servirá para interpretar la mayoría de resultados.

Cuando la SIG es > 0.05	No rechazar hipótesis nula	✓
Cuando la SIG es ≤ 0.05	Rechazar hipótesis nula	✗

La hipótesis nula de esta prueba es que la proporción del grupo 1 es igual al valor de prueba. El grupo 1 es, en todos los casos, aquel cuya codificación es la menor de los dos valores posibles (para este caso mujer), y el valor de prueba aquí es el valor por defecto 0.5. En el caso de la tabla 5, al leer la significancia (0.547), vemos que no rechazamos la hipótesis nula; es decir, la posibilidad de que los hombres tengan una proporción igual al de las mujeres (50% en cada uno).

Con la prueba binomial Melissa ya puede complementar la primera parte en la que solo mostraba que había más mujeres que varones. Ahora puede decir que proporcionalmente ambos grupos son probablemente iguales en la población, al ser ambos el 50%.

En esta prueba también podemos cambiar los porcentajes, y a continuación probaremos la probabilidad de distribución en la misma variable (sexo) pero con otros valores (Tabla 6): que la proporción de las mujeres en nuestra muestra sea de 80% (de hombres el 20%).

En este caso nuestros resultados nos indican que la significancia es menor a 0.05, por lo que la hipótesis nula es muy improbable, por esto la rechazamos.

Tabla 6. Prueba binomial de la variable sexo (proporción de prueba 0.80).

		Categoría	N	Proporción observada	Proporción de prueba	Significancia
Sexo	Grupo 1	Mujer	53	0.5	0.8	,000
	Grupo 2	Hombre	46	0.5		
	Total		99	1.00		

Elaboración propia.

Melissa refuerza la idea anterior al proponer que por más de que haya 7 mujeres más que los hombres en la muestra, esta diferencia no la vuelve una proporción que equivalga el 80% de la población. De la misma forma prueba con el 70% y el 60% y añade sus resultados a su informe. Guíate de los videos y prueba con otras proporciones tú también.

4.1.2. PRUEBA CHI-CUADRADO

Esta prueba la usaremos también cuando tengamos variables categóricas, pero en este caso ya pueden ser politómicas (con más de dos categorías). Para nuestro ejemplo, tomaremos la variable 'Asistencia al semestre 2010-1' e hipotetizaremos que la proporciones son iguales en los cuatro grupos (Tabla 7).

Tabla 7. Prueba chi-cuadrado de la variable asistencia al primer semestre.

Asistencia al semestre 2010-1

	N observado	N esperado	Residual
• Grupo 1: hasta el 80% de inasistencias	5	24.8	-19.8
• Grupo 2: hasta el 50% de inasistencias	26	24.8	1.3
• Grupo 3: hasta el 20% de inasistencias	61	24.8	36.3
• Grupo 4: hasta el 3% de inasistencias	7	24.8	-17.8
Total	99		

Estadísticos de contraste

	Asistencia al semestre 2010-1
• Chi-cuadrado	81,646a
• Gl	3
• Significancia asintótica	0.000

Elaboración propia.

Así, con la significancia obtenida (0.05) debemos rechazar la hipótesis que las proporciones son iguales para los cuatro grupos (25%). Es importante recordar que estas pruebas no nos dicen cuál es la distribución o proporción adecuada, solo prueban la hipótesis que ya viene por defecto en los programas.[20]

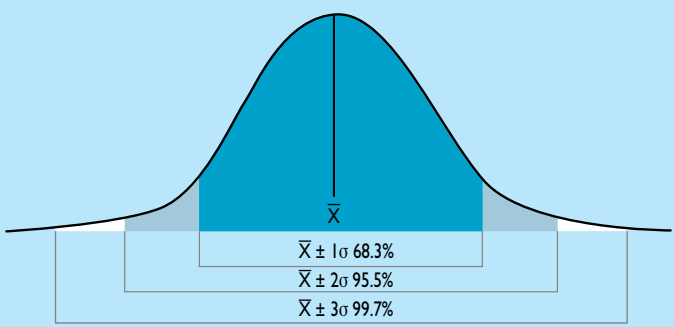
[20] Los valores por defecto son personalizables.

Tal como sugiere el resultado, Melissa comprueba que la distribución de las asistencias no es uniforme. Lo mismo equivale a decir que cada grupo no representa el 25% de todos los casos. Ella redacta que en su mayoría los estudiantes no son tan tardones contando a los que se encuentran en 'hasta el 3% de inasistencias' y los que están en 'hasta el 20% de inasistencias'. Ella realiza la prueba chi-cuadrado para las otras tres variables sobre la asistencia y puede mostrar si los porcentajes de asistencia crecen o decrecen. Anímate a probar si en algún semestre los tres grupos tuvieron el mismo porcentaje de casos.

4.2. PARA DATOS NUMÉRICOS

Antes de comenzar a trabajar con las variables numéricas debemos tener en claro el concepto de normalidad que será clave en nuestro análisis. Como vimos en el primer capítulo una curva muy importante es la 'normal'. Este tipo de curva indica que la distribución es simétrica, mesocúrtica y unimodal, formando la también llamada campana de Gauss. Esta curva es un modelo de distribución que muchas pruebas estadísticas exigen; es decir, se aplican a variables que deberían mostrar este comportamiento 'gaussiano' (Gráfico 8).

Gráfico 8. Distribución normal.



Elaborado por Jesús García.

Para ver si una variable tiene este comportamiento 'normal' debemos aplicar una prueba de normalidad, conocida como Shapiro-Wilk u otra llamada Kolmogorov-Smirnov, las cuales hipotetizan que la variable en cuestión se distribuye normalmente. Para la variable 'notas en ética' se aplicó esta prueba y se obtuvo (Tabla 8).

Tabla 8. Prueba de normalidad de la variable 'notas en ética'.

Pruebas de normalidad

	Kolmogorov-Smirnov			Shapiro-Wilk		
	Estadístico	gl	Significancia	Estadístico	gl	Significancia
Ética	0.119	97	0.002	0.963	97	0.008

Elaboración propia.

En la Tabla 8 podemos apreciar que la variable 'notas en ética' no sería considerada normal, ya que por su significancia en ambas pruebas (0.002 o 0.008) se rechaza la hipótesis nula que proponía que esta variable se distribuía normalmente.

Es importante recalcar que hacemos estas diferencias pues cuando una variable es normal podrá ser mejor tratada con técnicas paramétricas a diferencia de las que no pasan la prueba de normalidad en cuyo caso se puede optar por técnicas no paramétricas.

Melissa había sacado el histograma de la variable 'ética' para la parte de exploración y pudo ver que no era normal, pero no sabía cómo probarlo. Luego de que aprendiera la prueba de Kolmogorov-Smirnov pudo añadir a su análisis que la moda de la variable verdaderamente estaría impidiendo la formación de una curva normal.

A continuación pasaremos a explicar las 'pruebas T' en las que realizaremos el análisis univariado para variables numéricas, no debemos olvidar que este tipo de pruebas asumen la normalidad; es decir, este es un requisito previo.

4.2.1. PRUEBA T PARA UNA MUESTRA

Esta prueba permite saber dos cosas. Primero, cuál es el intervalo de confianza de la media de la población con la media muestral conocida, y segundo probar si algún valor en particular puede ser aceptado como media poblacional posible.

La salida es la misma en todos lo casos y se ve en la Tabla 9.

Tabla 9. Prueba T (para una muestra).

Estadísticos para una muestra

	N	Media	Desviación típica	Error típico de la media
Matemática I	99	14.08	3.431	0.345

Prueba para una muestra

	Valor de prueba = 0					
	t	gl	Significancia(bilateral)	Diferencia de medias	95% Intervalo de confianza para la diferencia	
					Inferior	Superior
Matemática I	40.839	98	0.000	14.081	13.40	14.77

Este resultado muestra 'valor de prueba = 0'; cuando eso es así, lo que estamos buscando es saber, cual es el intervalo de la media poblacional al 95% de confianza sabiendo que la media de la muestra es 14.08. Vemos en los recuadros sombreados que la media poblacional puede estar entre 13.40 y 14.77, pero como es un intervalo al 95% de confianza, hay 5% de probabilidad que no esté ahí.[21]

Si el valor de prueba no fuese 'cero', lo que se hipotetiza es que ese valor de prueba es la media poblacional. De nuevo, si la significancia fuese menor o igual a 0.05 se rechazaría esa hipótesis.

[21] El valor de 95% es el valor que los programas usan por defecto en general, pero eso se puede alterar según desee el investigador.

Entusiasmada con la parte inferencial de la estadística, Melissa quiere aplicar la prueba T para una muestra de todas las variables numéricas de su data. Quiere ver si en todo Estudios Generales Letras, con un 95% de confianza, hay posibilidad de que en todo el alumnado salga aprobado. Ella ha encontrado resultados interesantes conforme se avanza en el tiempo como el hecho de que los alumnos en general no cambian su rendimiento, ni para bien ni para mal. Melissa apunta que las motivaciones como la de acabar bien Letras o tener buen CRAEST no parecerían influir en el desempeño general por curso.

4.2.2. PRUEBAS ALTERNATIVAS

Cuando no se cumplen los requisitos de la prueba t se tiene alternativas no paramétricas. Por ejemplo, tenemos la prueba Wilcoxon que se utiliza para variables ordinales y numéricas tanto en el R como en el SPSS 19. Para el caso descrito de la prueba T, la prueba Wilcoxon hace exactamente lo mismo, tiene las mismas hipótesis y los resultados se interpretan de la misma manera.

Para las variables que no cumplían con la normalidad, Melissa optó por la prueba Wilcoxon y así terminó de hacer su análisis sobre el rendimiento del alumnado por curso. Dado que esta opción en el SPSS no nos ofrece el intervalo para la media, ella trató de poner valores cercanos a la media y así sugerir los que salían no significativos. Para acelerar el proceso eligió usar la sintaxis para evitar hacer varios clicks.

El cuadro de diálogo de la prueba 'binomial' del SPSS permite hacer inferencias sobre la las medidas de posición; por ejemplo, permite que hipoteticemos si algún valor de prueba puede ser la mediana. En la Tabla 10 vemos si es probable que la mediana de la 'escala económica en el cuarto semestre' sea 3.

Tabla 10. Prueba binomial para la variable 'escala económica en el cuarto semestre' (valor de prueba = 3).

Prueba binomial

		Categoría	N	Proporción observada	Proporción de prueba	Significancia
Escala económica en el cuarto semestre	Grupo 1	≤ 3	88	0.89	0.50	
	Grupo 2	> 3	11	0.11		,000a
	Total		99	1.00		

Elaboración propia

Como podemos ver en las tablas al poner como valor de prueba 3 se rechaza la hipótesis a diferencia de la Tabla 11 en la que el valor de prueba es 2. Esto quiere decir que lo más probable es que en la población la mediana de la escala económica en el cuarto semestre sea 2.

Tabla 11. Prueba binomial para la variable 'escala económica en el cuarto semestre' (valor de prueba = 2).

Prueba binomial

		Categoría	N	Proporción observada	Proporción de prueba	Significancia
Escala económica en el cuarto semestre	Grupo 1	≤ 2	46	0.46	0.50	
	Grupo 2	> 2	53	0.54		,547[a]
	Total		99	1.00		

Elaboración propia.

Melissa hizo la prueba binomial para las variables ordinales de su data, tanto para las que preguntaban sobre la asistencia como para las que preguntaban sobre la escala económica de todo el alumnado. Así como con el caso de la media no encontró mucha variación conforme se avanzaba en el tiempo.

5. INFERENCIA BIVARIADA

La inferencia bivariada, como su nombre lo indica, analiza la relación existente entre dos variables. Al igual que en la sección anterior, presentaremos las relaciones según tipos de escala que intervengan.

5.1. RELACIÓN NUMÉRICA-NUMÉRICA

Cuando se analizan dos variables numéricas podemos plantearnos dos escenarios:

– el primero (y más difundido) es cuando se analizan dos variables diferentes en un corte de tiempo;

– el segundo caso es cuando se analiza una misma variable medida en dos momentos diferentes.

El primer caso es conocido como correlación y se puede utilizar dos pruebas que se diferencian por sus requisitos.

– En el caso de la primera alternativa se presenta la técnica paramétrica[22] conocida como la R de Pearson,

– y la otra alternativa, la Rho de Spearman (no paramétrica).

El segundo caso, conocido como diferencia de medias para muestras relacionadas, tiene también dos pruebas alternativas.

– La primera es la prueba T (para el caso paramétrico)

– y la segunda la prueba Wilcoxon (caso no paramétrico).

5.1.1. R DE PEARSON

Lo que se analiza en esta prueba es la existencia o no de la correlación lineal entre dos variables numéricas que se distribuyan normalmente.[23]

[22] Este tipo de pruebas, en este caso, tienen como requisito la normalidad y la igualdad de varianzas.
[23] Nos referimos a la curva normal (Gauss).

Tabla 12. Correlación lineal (R de Pearson).

Correlaciones

		Matemática 1	Matemática 2
• Matemática 1	Correlación de Pearson	1	,701**
	Significancia (bilateral)		0.000
	N	99	97
• Matemática 2	Correlación de Pearson	,701**	1
	Significancia (bilateral)	0.000	
	N	97	97

** La correlación es significativa al nivel 0,01 (bilateral).

Elaboración propia.

En el ejemplo de la Tabla 12 se busca saber si había o no correlación entre las notas de Matemática 1 y Matemática 2. En este caso la hipótesis base de la prueba de Pearson es que 'no hay correlación'[24]. Por el resultado de la significancia obtenida (0.00) rechazaremos la hipótesis de la no correlación, concluyendo que sí hay correlación. Luego vemos el valor del estadístico que ha resultado: 0.701, que indica que la correlación entre las notas en Matemática 1 y Matemática 2 es alta y directa.[25]

Un caso particular de correlación es la correlación parcial. Esta prueba, nos permite verificar que la correlación hallada no sea espuria al eliminar el efecto de una tercera variable entre la relación de otras dos.

En el siguiente ejemplo (Tabla 13) trataremos de corroborar que sí hay correlación entre Matemática 1 y Matemática 2, controlando dicha prueba con la variable edad. En esta prueba llamada correlación parcial buscamos demostrar que, efectivamente, nuestras variables Matemática 1 y Matemática 2 se correlacionan aunque quitemos el efecto de la variable edad:

[24] Es decir, que el r de Pearson es igual a 0.

[25] El r de Pearson oscila entre -1 y 1. Cuando es cero no hay correlación, cuando es positivo la correlación es inversa (cuando una variable aumenta la otra disminuye, o viceversa) y cuando es positivo es directo (ambas aumentan o disminuyen). Valores superiores a 0.7 o -0.7 se consideran altos, menores a 0.4 se consideran bajos.

Tabla 13. Correlación parcial.

Variables de control			Matemática 1	Matemática 2
Edad	Matemática 1	Correlación	1.000	0.707
		Significancia (bilateral)	.	0.000
		Gl	0	94
	Matemática 2	Correlación	0.707	1.000
		Significancia (bilateral)	0.000	.
		Gl	94	0

Elaboración propia.

De los resultados obtenidos debemos observar la significancia obtenida en la parte sombreada de la tabla (ya que es la tiene el resultado controlado por 'edad'). Como vemos la significancia es 0.00 y el r es 0.707, por lo que podemos decir que entre nuestras variables Matemática 1 y Matemática 2 existe una intensa correlación directa habiendo controlado la variable edad.

Si bien las pruebas de inferencia univariada le parecieron interesantes, más aun fueron las de bivariada. Establecer relaciones le fue muy útil a Melissa ya que pudo responder a las preguntas que tenía en un inicio. La correlación lineal de Pearson le sirvió para buscar relación entre las notas, el examen de admisión, el examen de admitidos y la edad. Tal como el caso mostrado, Melissa encontró que los cursos de filosofía se correlacionaban entre ellos, así como los de historia, pero no entre estos dos grupos de cursos. Anímate a probar si puede existir correlación entre la edad y el examen de admisión o cualquier otro curso. No te olvides de guiarte con los videos.

5.1.2. Rho de Spearman

En breve, el Spearman es la versión no paramétrica disponible para saber si hay correlación entre dos variables numéricas. Su interpretación es idéntica al r de Pearson (ver pie de página 24), pero no exige que las variables se distribuyan normalmente ni que la relación entre ellas sea lineal. En Tabla 14, hemos utilizado para el ejemplo la variable 'notas en Ética' y 'notas en Matemática 2'.

Tabla 14. Correlación Spearman.

Correlaciones

			Ética	Matemática 2
Rho de Spearman	Ética	Coeficiente de correlación	1.000	-.038
		Significancia (bilateral)	.	0.713
		N	97	95
	Matemática 2	Coeficiente de correlación	-.038	1.000
		Significancia(bilateral)	0.713	.
		N	95	97

Elaboración propia.

En el ejemplo anterior buscamos ver si existe correlación entre las notas de 'Ética' y 'Matemática 2', al leer la significancia bilateral aceptamos la hipótesis 'no hay correlación' (ya que es mayor a 0.05), por lo que podemos decir que no hay correlación entre las notas de 'Ética' y de 'Matemática 2'.

Dado que son no pocas variables las que pasan la prueba de normalidad, requisito para la prueba de correlación de Pearson, Melissa tuvo que ampararse en la versión no paramétrica, en la correlación de Spearman. Los resultados, como en la mayoría de los casos, no se contradicen pero es necesario tener en mente los requisitos. La correlación más fuerte que ella ha encontrado es entre las notas de Ecología y Derecho con un coeficiente de Spearman de 0.906 lo que le hizo afirmar que los que salieron bien en Ecología el último ciclo de Estudios Generales Letras también salieron bien en Derecho. Sin embargo, por ser cursos de materias diferentes no se animó a buscar una posible explicación. Si tú llevaste ambos cursos en Estudios Generales Letras, ¿podrías proponer alguna? Asimismo, anímate a buscar una correlación fuerte como la de Melissa.

5.1.3. Prueba T para muestras relacionadas

La r de Pearson analizaba la correlación entre dos diferentes variables, pero la prueba T analiza la relación entre la misma variable medida en dos oportunidades diferentes a las mismas unidades de estudio en una muestra. La prueba T se encarga de informarnos si las dos medias de estos dos momentos son probablemente iguales o no. Los resultados para esta prueba se ven en la Tabla 15.

Tabla 15. Prueba T (muestras relacionadas).

Prueba de muestras relacionadas

		Diferencias relacionadas					T	gl	Significancia (bilateral)
		Media	Desviación típica	Error típico de la media	95% intervalo de confianza para la diferencia				
					Inferior	Superior			
Par 1	Examen de Admisión - Examen de Admitidos	-60.081	61.148	6.146	-72.277	-47.885	-9.776	98	.000

Elaboración propia

En este caso, tenemos la hipótesis 'no hay diferencia de valores medios' y como la significancia es menor a 0.05 rechazamos la hipótesis y concluimos que lo más probable es que los promedios no sean los mismos; es decir, que hay un cambio de un examen a otro en la misma población.

Para el caso de diferencias, Melissa optó por analizar la relación existente el puntaje obtenido en el examen de admisión y el examen de admitidos, ambos vistos como exámenes similares al ser exámenes para ingresar a Estudios Generales Letras. Quería ver si los valores medios de ambas variables se diferenciaban y lo que encontró fue que no había igualdad. En la data de Melissa, a su vez, no se encuentran otras variables que se hayan medido dos veces dado que los cursos son solo una vez y las otras variables son categóricas.

5.2. RELACIÓN CATEGÓRICA-CATEGÓRICA

5.2.1. Chi-cuadrado de Pearson

Cuando estudiamos dos variables de tipo categóricas queremos ver si existe o no asociación entre dos variables, la prueba que se utiliza para estos fines es la del chi-cuadrado, esta prueba estadística nos permite hallar la relación para este tipo de variables. Esta prueba solo permite saber si una variable categórica está relacionada con otra de la misma escala.

El siguiente ejemplo (Tabla 16) informa si hay o no asociación entre las variables 'asistencia en el primer semestre' (asistencia 1) y 'escala en el primer semestre' (escala 1) en el primer semestre (en este caso ambas son ordinales).

Tabla 16. Asociación entre dos ordinales.

Pruebas de chi-cuadrado

	Valor	gl	Significancia asintótica (bilateral)
• Chi-cuadrado de Pearson	29,752[a]	12	.003
• Razón de verisimilitudes	31.406	12	.002
• Asociación lineal por lineal	16.970	1	.000
• N de casos válidos	99		

[a] 13 casillas (65.0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es .56.

Elaboración propia.

En la tabla obtenida debemos leer la fila 'chi-cuadrado de Pearson' y revisar la significancia (valor sombreado). La hipótesis nula aquí es 'ambas variables son independiente' (no están asociadas); y dado que la significancia es 0.003 rechazaremos la hipótesis, concluyendo que es muy probable que haya asociación entre las variables 'asistencia 1' y 'escala 1'.

Debemos saber que esta prueba no proporciona información acerca del sentido de la relación (directo o inverso) ni sobre la intensidad de la misma; por ello, el chi-cuadrado requiere de análisis complementarios para conocer ambas cosas. Las pruebas estadísticas para este tipo de pruebas varían de acuerdo a la escala de la variable categórica.

5.2.2. PRUEBAS PARA NOMINALES: SOLO INTENSIDAD

Cuando al menos una de las dos variables categóricas a analizar es nominal, se tiene disponibles las siguientes pruebas:

- Coeficiente de Contingencia
- Phi (para tablas 2x2)
- V de Cramer
- Lambda
- Coeficiente de Incertidumbre

Las medidas Lambda y Coeficiente de incertidumbre son utilizadas cuando la hipótesis ubica a cada variable categórica en el rol de independiente y dependiente. A estas medidas se les denomina de tipo direccional.[26] En todos estos casos, el resultado está entre 0 y 1,[27] y su interpretación se basa en la Tabla 17.

Tabla 17. Interpretación de la asociación.

• Si el coeficiente es menor que 0,400	La relación es despreciable.
• Si el coeficiente está entre 0,400 y 0,600	La relación es medianamente fuerte.
• Si el coeficiente es mayor que 0,600	La relación es fuerte.

5.2.3 PRUEBAS PARA ORDINALES: INTENSIDAD Y SENTIDO

Cuando las dos variables son ordinales se tienen las siguientes pruebas:

- Tau-B (tablas cuadradas – nxn)
- Tau–C
- Gamma
- D de Sommers

La D de Sommers es la única medida direccional de este grupo. En todos los casos las medias van de -1 a 1, y su interpretación es similar al r de Pearson (ver pie de página 28).

[26] A las otras se les denomina de tipo simétrico.
[27] Algunas no llegan a 1, pero en general es el valor de referencia.

A Melissa le fue muy útil el chi-cuadrado porque su data tiene varias variables categóricas. En un comienzo, ella optó por entablar varias relaciones entre las variables pero le faltaba el estadístico que le permitiera decir si la relación era estadísticamente significativa. A su vez las pruebas auxiliares al chi-cuadrado le dieron una idea de la intensidad y del sentido de la relación (si eran ordinales). Además del ejemplo anterior, ella encontró que la variable 'Escala económica en el primer semestre' se asociaba significativamente con 'Nivel del Inglés'. Dado que esta última tiene una codificación inversa a la anterior y el estadístico de Gamma le salió -0.98, ella interpretó que 'a mayor ubicación en la escala económica de la PUCP, el nivel del inglés es mejor'.

Así como ella, anímate a practicar entablando relaciones entre las variables categóricas. Melissa hizo lo mismo e indagó más sobre el uso del chi-cuadrado y las medidas auxiliares. Ella encontró que las medidas direccionales que ofrece el SPSS no son tan robustas, y que en su lugar podría usar las pruebas de regresión, tema que no vamos a ver en este manual por tener un grado de dificultad mayor. Pero las medidas simétricas sí son muy útiles ya que funcionan como leer los coeficientes de correlación tanto de Pearson como de Spearman.

5.3. RELACIÓN CATEGÓRICA-TUMÉRICA

5.3.1. PRUEBA T PARA MUESTRAS INDEPENDIENTES

La prueba T para muestras independientes requiere dos variables: una dependiente (que será numérica) e independiente (que será categórica dicotómica). La Tabla 18 muestra la prueba de hipótesis que la media del examen de admisión ha sido el mismo para hombres y mujeres.

Lo que se busca en la Tabla 18 es saber si hay alguna diferencia en la notas obtenidas en el examen de admisión en función de la variable sexo. En este caso, la hipótesis base de la prueba es 'no hay diferencia de valores medios'; es decir, que el factor usado no tuvo efecto. En este caso, al leer la significancia aceptamos la hipótesis y, por lo tanto, concluimos con que el factor sexo no ha causado diferencias en las notas del examen de admisión.

Tabla 18. Prueba T para muestras independientes.

Estadísticos de grupo

	Sexo	N	Media	Desviación típica	Error típico de la media
Examen de Admisión	Mujer	53	653.34	111.584	15.327
	Hombre	46	650.35	102.615	15.130

Prueba de muestras independientes

		Prueba de Levene para la igualdad de varianza		Prueba T para la igualdad de medias						
		F	Significancia	t	gl	Significancia (bilateral)	Diferencia de medias	Error típico de la diferencia	95% intervalo de confianza para la diferencia	
									Inferior	Superior
Examen de Admisión	Se han asumido varianzas iguales	.541	.464	.138	97	.890	2.992	21.666	-40.009	45.992
	No se han asumido varianzas iguales			.139	96.659	.890	2.992	21.537	-39.755	45.738

Elaboración propia.

La prueba T para muestras independientes le fue útil a Melissa cuando tuvo que ver si había diferencia entre las dos categorías de una variable dicotómica. Este fue el caso cuando quiso ver si había diferencia entre las notas obtenidas según la variable 'sexo'. En lo particular, ella creía que no existía diferencia significativa entre las medias de las notas de los hombres de los de las mujeres. Esto porque en las entrevistas que ella había hecho cuando ingresaron a Estudios Generales Letras vio que en ambos grupos se presentaban las mismas capacidades analíticas y tenían habilidades parecidas. Anímate a hacer algunas pruebas T usando las variables que preguntan sobre si le gustó la universidad o no, o si se cambiaría a Estudios Generales Ciencias o no, como dicotómicas.

Esta prueba requiere que previamente se verifique si se cumple el supuesto de igualdad de varianzas, para lo cual se utiliza la prueba de Levene que aparece a la izquierda de la tabla. Aquí la significancia ha salido 0.464, por lo que no se rechaza la hipótesis de la prueba de igualdad de varianzas, por lo que se procede a interpretar la primera fila. Si hubiera salido menor o igual a 0.05, se hubiera tenido que reportar la significancia de la fila de abajo.

5.3.2. PRUEBA F (ANOVA DE UN FACTOR)

Cuando deseamos conocer si existe efecto de un factor (variable categórica) en los valores medios de una variable numérica recurrimos al ANOVA de un factor. Esta técnica es más compleja, pues es parte del modelo lineal general univariado y tiene como hipótesis nula que: 'no hay efecto del factor en los valores medios de la variable dependiente'. Por lo tanto, lo que se obtiene con esta prueba es saber si hay una diferencia de los promedios de una variable (numérica) en más de dos grupos -delimitados por nuestra variable categórica-. Por ejemplo, haríamos uso de esta técnica si queremos conocer si existe efecto de la religión de una persona (católico, protestante, judío, etc.) en su nivel de ingresos.

En el siguiente ejemplo tenemos como variable numérica dependiente la variable 'Matemática 2' y como variable independiente la categórica 'asistencia al primer semestre' (asistencia 1). Queremos saber si la asistencia en ese semestre tiene algún efecto sobre las notas en matemática 2. Para ello debemos seguir una secuencia:

Primero debemos verificar si hay homogeneidad de varianzas. Para ello pedimos la prueba de Levene, resultados que se presentan en el Gráfico 9. Dado que la hipótesis nula de la prueba es que hay homogeneidad de varianzas, la significancia de 0.018 nos indica que debemos echar esa hipótesis.

Gráfico 9. Prueba de homogeneidad de varianzas.

Prueba de homogeneidad de varianzas

Matemática 2			
Estadístico de Levene	gl1	gl2	Significancia
• 3.526	3	93	0.018

La prueba F del ANOVA puede no ser adecuada cuando no hay homogeneidad de varianzas. En este caso, al ver los resultados de la Tabla 19: Tabla ANOVA, concluimos que las medias de Matemáticas 2 no son iguales, pues la significancia es 0.000, por lo que la hipótesis nula debe ser rechazada.

Tabla 19. Tabla ANOVA.

ANOVA

Matemática 2					
	Suma de cuadrados	gl	Media cuadrática	F	Sig.
• Inter-grupos	338.490	3	112.830	16.034	0.000
• Intra-grupos	654.417	93	7.037		
Total	992.907	96			

Elaboración propia.

Cuando esto sucede; es decir, cuando no se puede asegurar homogeneidad de varianzas se suele solicitar 'Pruebas robustas de igualdad de las medias' (ver Tabla 20).

Tabla 20. Pruebas robustas del ANOVA.

Matemática 2				
	Estadístico[a]	gl1	gl2	Significancia
• Welch	58.897	3	21.548	0.000
• Brown-Forsythe	20.999	3	19.988	0.000

[a] Distribuidos en F asintóticamente.
Elaboración propia.

Leeremos la tabla de pruebas robustas puesto que ellas se utilizan cuando no hay homogeneidad de varianzas. Para la lectura de resultados tenemos dos opciones:

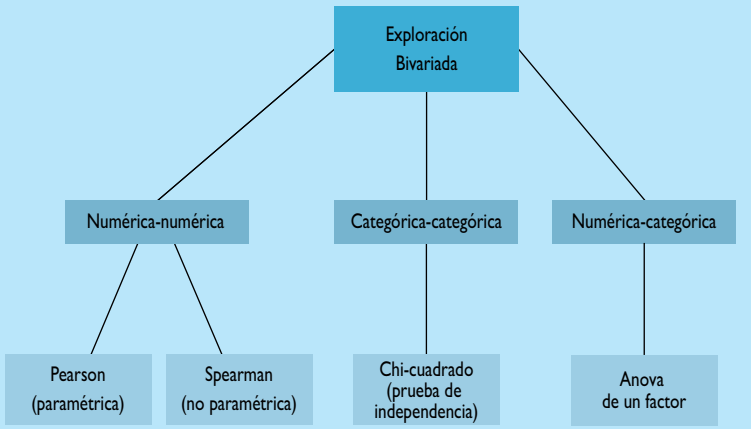
– Welch: se usa cuando hay igualdad en el número de casos de cada una de las modalidades de la variable independiente categórica.
– Brown-Forsythe: se usa cuando hay diferente número de casos en cada una de las modalidades de la variable independiente categórica.

La prueba ANOVA fue la última que utilizó Melissa en su informe. Esta prueba por permitir el uso de dos grupos o más, bien podría sustituir a la prueba T. Aunque en su formato original necesita que la distribución de los casos sea homogénea, los estadísticos auxiliares nos permiten pasar este requisito y habría que enfocarse si hay una misma cantidad de grupos o no. Ella usó esta prueba con todas las variables categóricas politómicas de su data. La que más le sorprendió fue el factor 'nivel de inglés' ya que si bien no es muy exigente el uso de inglés en los cursos de Letras, sí marca la diferencia conocer este idioma en algunos cursos como Ecología y Derecho.

En este caso, leemos Welch puesto que hay diferente número de casos en las modalidades de la variable asistencia en el primer semestre. La significancia de este estadístico nos dice que debemos rechazar la hipótesis de igualdad de medias; por lo que concluimos con mayor seguridad que las medias son diferentes y que efectivamente hay un efecto de nuestro factor (Asistencia en el primer semestre) sobre nuestra variable dependiente (Notas de Matemática 2).

RESUMEN

Gráfico 10. Resumen Inferencia bivariada.



Elaboración propia

Melissa entregó su informe en la última semana del tercer mes. Adjuntó a las tablas de frecuencias, los estadísticos y los gráficos, un informe de carácter más cualitativo sobre los alumnos. Ella llegó entrevistar a los 99 alumnos que fueron parte de su data y pudo entrar en detalle sobre los motivos por qué faltaban a clases o por qué no habían obtenido buen puntaje en determinados cursos. Encontró con esta investigación que influye mucho la carrera que el alumno o alumna quiera seguir, ya que le ponen mayor empeño a los cursos afines o estudian más aquellos cursos que les servirán de base para aprender otros más avanzados cuando ingresen a sus facultades. Melissa sugirió consignar esta pregunta la próxima vez que alguien armara una data parecida a los que revisarían su informe. Pasada una semana le dijeron que su trabajo era muy bueno y que deseaban contar con ella durante todo un año, pero ella con la modestia del caso rechazó la oferta. Se dio cuenta que podía aplicar lo que aprendió de estadística en una investigación que le llamara más la atención, iba a utilizar la estadística en su tesis de Licenciatura.

Ejercicios 2

1. **Los que sacaron buena nota en el curso de Teoría General del Lenguaje también sacaron buena nota en el curso de Redacción y Argumentación.**
   a. Falso
   b. Verdadero
2. **El gusto de la carrera en el primer y segundo año no se asocia con que sea hombre o mujer.**
   a. Verdadero
   b. Falso
3. **Existe la creencia de que las mujeres asisten con más frecuencia a clases a comienzos del segundo año.**
   a. Falso
   b. Verdadero
4. **Marque si es verdadero:**
   a. Los hombres tuvieron más puntaje que las mujeres en el examen de admisión.
   b. Los puntajes en el examen de admisión son diferentes según la escala.
   c. A los que les gustó más el primer año tuvieron más notas que a los que no les gustó la carrera.
   d. No hay diferencia entre los que quieren cambiarse y no a Estudios Generales Ciencias según las notas del curso Historia antigua y medieval.
5. **Marque la opción incorrecta, puede marcar más de una:**
   a. Hay diferencia en las notas de Fe y Cultura Actual según la escala de pago del tercer semestre.

b. La escala de pago que estaría originando la diferencia es el 'quinto'.

c. Cuando se tenga homogeneidad de varianzas tenemos que leer la prueba T2 de Tamhane.

6. Marque la alternativa correcta:

a. No hay diferencia entre el examen de admisión y el examen de admitidos.

b. No hay diferencia en el puntaje del examen de entrada según el nivel de inglés.

c. Hay diferencia en el examen de admitidos según la escala socioeconómica del primer semestre.

7. Marque la alternativa incorrecta, puede marcar más de una:

a. Hay correlación positiva fuerte entre los 2 cursos de filosofía de Letras y ética.

b. La correlación anterior desaparece si controlamos por la edad de los alumnos.

c. Hay correlación negativa débil entre los cursos de matemática 1 y de el Perú en la Historia de América.

8. Marque la alternativa correcta:

a. Las notas de Investigación Académica y Quechua se distribuyen de una manera normal.

b. Las mujeres representan el 80% de la población.

c. Los que quieren cambiarse a Ciencias en el primer año representan el 10%.

d. La distribución no es uniforme en la variable nivel de inglés.

9. Marque la respuesta incorrecta, puede marcar más de una.

a. La asistencia del primer es diferente a los otros tres ciclos a nivel de la población de todo Letras.

b. En la población de Letras 60% se encuentra en el nivel avanzado, 30% en el nivel intermedio y 10% en nivel básico del inglés.

c. Los que se quieren cambiar en el segundo año representan el 15% de la población.

SOLUCIONARIO

EJERCICIOS 1

1. ¿Cuál de los cursos que se llevaron el primer ciclo tiene más puntaje en promedio?

   c. Historia antigua y medieval

2. Marque si es correcto:

   a. La media del curso de matemática 1 es 14.
   Correcto, sin contar los decimales.

   b. La nota que más se repite es 16.
   Incorrecto, la nota que más se repite es 13.

   c. La mediana es la misma en los 3 cursos de historia de letras.
   Correcto, sin contar decimales.

   d. Existe mucha variación en los cursos de matemática.
   Incorrecto, en los dos cursos de matemática no hay mucha variación.

   e. Hay más alumnos en las dos primeras escalas del primer semestre.
   Incorrecto, hay más alumnos en las otras tres escalas restantes.

   f. Hay más desaprobados en los dos cursos de matemática que en el resto de cursos.
   Correcto, son los dos cursos que más desaprobados tienen.

   g. El curso de apreciación musical es el que agrupa las mejores calificaciones en el semestre.
   Correcto, contando las frecuencias de las notas 18, 19 y 20, apreciación musical concentra las mejores notas.

   h. El curso de lógica es el que tiene más desviación típica tiene.
   Incorrecto, matemática 1 y matemática 2 son los que más dispersión tienen.

3. ¿Qué cursos se distribuyen de manera normal en el primer ciclo?

   Ninguno. De acuerdo a la prueba de normalidad de Kolmogorov-Smirnov, ningún curso en el primer semestre tiene una distribución normal. Usar las prueba que se encuentra en el menú explorar, por la corrección de Lilliefors para muestras pequeñas.

4. ¿Cuántos cursos del segundo semestre son los que tienen una distribución sesgada?

   Cuatro. Los cursos de Filosofía Moderna, Lógica y epistemología, Elementos de Ciencia Política y El Perú en la historia de América son los que tienen una distribución más sesgada. Esto se puede comprobar mirando el estadístico de asimetría y los histogramas de las variables.

5. Se suele decir que:
   a. Los cursos de matemática tienen cada uno un 30% de alumnos desaprobados.
      Falso. En Matemática 1 hay un 18% de desaprobados y en Matemática 2, un 12%.
   b. El curso de Apreciación Musical es el que tiene más gente aprobada.
      Falso. Hay varios cursos como el de Elementos de Ciencia Política que no tiene ningún desaprobado.
   c. Hay más desaprobados en el primer ciclo comparado con el segundo.
      Verdadero. En el primer ciclo hay un 30% de desaprobados y en el segundo, un 13%.
6. El coeficiente de variación y rango son medidas de:

   Respuesta a. Dispersión.
7. Con el boxplot podemos ver la …….de la variable ……..

   Respuesta b. Posición, simetría, dispersión – ordinal y numérica.
8. La barra de error con la desviación estándar es una gráfica que:

   Respuesta b. Visualiza la dispersión de la media de la variable
9. El histograma permite ver:

   Respuestas a, b, c y d.
   a. El apuntamiento de la variable numérica discreta.
   b. La simetría y apuntamiento de la variable numérica continua.
   c. La concentración y dispersión de la variable numérica discreta.
   d. La concentración de la variable numérica discreta.
10. El boxplot no muestra:

    Respuestas a, b y c
    a. El rango intercuartílico de la variable ordinal.
    b. La simetría de la variable numérica.
    c. La mediana como medida de centralización.
11. De las medidas de centralización:

    Respuestas a, b, c y d.
    a. La media es la medida más sensible a los valores extremos.
    b. La mediana es la medida menos sensible a los valores atípicos.
    c. La moda es una medida tanto para variables numéricas y categóricas
    d. La media pierde representatividad si la distribución es sesgada.

EJERCICIOS 2

1. Los que sacaron buena nota en el curso de Teoría General del Lenguaje también sacaron buena nota en el curso de Redacción y Argumentación.

   Respuesta a. Falso. La correlación de Pearson no es significativa y no podemos sostener tal afirmación.

2. El gusto de la carrera en el primer y segundo año no se asocia con que sea hombre o mujer.

   Respuesta a. Verdadero. No existe asociación entre las dos variables. El chi-cuadrado sale no significativo.

3. Existe la creencia de que las mujeres asisten con más frecuencia a clases a comienzos del segundo año.

   Respuesta a. Falso. Para nuestra muestra podemos afirmar el enunciado, pero no para toda la población de Letras. El chi-cuadrado sale no significativo

4. Marque si es verdadero:

   a. Los hombres tuvieron más puntaje que las mujeres en el examen de admisión.
      Falso. La prueba T para muestras independientes sale no significativa.
   b. Los puntajes en el examen de admisión son diferentes según la escala.
      Falso. La prueba ANOVA sale no significativa.
   c. A los que les gustó más el primer año tuvieron más notas que a los que no les gustó la carrera.
      No necesariamente. En varios cursos del primer año no hay diferencia significativa entre los dos grupos.
   d. No hay diferencia entre los que quieren cambiarse y no a Estudios Generales Ciencias según las notas del curso Historia antigua y medieval.
      Verdadero. La prueba T sale no significativa.

   Respuesta d.

5. Marque la opción incorrecta, puede marcar más de una:

   Respuestas a, b y c.

   a. Hay diferencia en las notas de Fe y Cultura Actual según la escala de pago del tercer semestre.
   b. La escala de pago que estaría originando la diferencia es el 'quinto'.
   c. Dado que tenemos homogeneidad de varianzas tenemos que leer la prueba T2 de Tamhane.

6. Marque la alternativa correcta:

   a. No hay diferencia entre el examen de admisión y el examen de admitidos.
      Falso. La prueba T para muestras relacionadas sale significativa.
   b. No hay diferencia en el puntaje del examen de entrada según el nivel de inglés.
      Verdadero. La prueba ANOVA sale no significativa.

c. Hay diferencia en el examen de admitidos según la escala socioeconómica del primer semestre.
Falso. La prueba ANOVA sale no significativa.

Respuesta b.

7. Marque la alternativa incorrecta:
   a. Hay correlación positiva fuerte entre los dos cursos de filosofía de Letras y ética.
   Incorrecta. Solo con Filosofía Moderna sale significativa y con Filosofía Antigua, no..
   b. La correlación anterior desaparece si controlamos por la edad de los alumnos.
   Incorrecta. La correlación entre Filosofía Moderna y Ética no desaparece.
   c. Hay correlación negativa débil entre el curso de matemática 1 y el Perú en la Historia de América.
   Incorrecta. No hay correlación significativa.

   Respuestas a, b, y c.

8. Marque la alternativa correcta:
   a. Las notas de Investigación Académica y Quechua se distribuyen de una manera normal.
   Falso. Según la prueba de normalidad de Kolmogorov-Smirnov, ninguno de los dos cursos tiene una distribución normal. Usar las prueba que se encuentra en el menú explorar, por la corrección de Lilliefors para muestras pequeñas.
   b. Las mujeres representan 80% de la población.
   Falso. La prueba binomial sale significativa.
   c. Los que quieren cambiarse a Ciencias en el primer año representan el 10%.
   Verdadero. La prueba binomial nos sale no significativa.
   d. La distribución no es uniforme en la variable nivel de inglés.
   Falso. La prueba chi-cuadrado para ver si la distribución es uniforme no sale significativa.

   Respuesta c.

9. Marque la respuesta incorrecta.

   Respuestas a, b. y c.

   a. La asistencia del primer es diferente a los otros tres ciclos a nivel de la población de todo Letras.
   Falso. Existe una asociación significativa entre las cuatro variables de asistencia y los estadísticos de simetría son positivos.
   b. En la población de Letras 60% se encuentra en el nivel avanzado, 30% en el nivel intermedio y 10% en nivel básico del inglés.
   Falso. La prueba de chi-cuadrado sale significativa.
   c. Los que se quieren cambiar en el segundo año representan el 15% de la población.
   Falso. La prueba chi-cuadrado sale significativa lo que indica que la proporción es menor.

LINKS DE LA GUÍA

- Mindmeister:
  http://www.mindmeister.com/107137229#

- Wiki:
  http://wiki.pucp.edu.pe/estadisticavirtual/

- VideosPUCP:
  http://videos.pucp.edu.pe/usuarios/administrar/3119