INSTITUTO SUPERIOR DEL PROFESORADO DE SALTA Nro. 6005

PLAN PEDAGÓGICO

- Llenar la propuesta en forma concreta, clara y útil para los estudiantes, las actividades deben estar acompañadas por los links en PDF de la bibliografía solicitada además de video, blogs, sitio web o mail para la realización y entrega de actividades.
- Una vez completo se debe enviar por mail al Coordinador de Carrera a la cual pertenezca.

INSTITUTO SUPERIOR DEL PROFESORADO DE SALTA Nro. 6005

PLAN PEDAGÓGICO – N° 3

CARRERA: TECNICATURA SUPERIOR EN ADMINISTRACIÓN CON ORIENTACIÓN EN COMERCIALIZACIÓN

(DESDE EL 05 DE OCTUBRE AL 16 OCTUBRE DEL 2020)

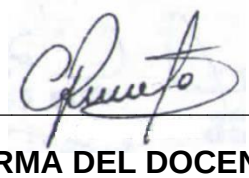
ASIGNATURA: ESTADÍSTICA

APELLIDO Y NOMBRE DEL DOCENTE: PINTO, CRISTIAN VICTOR

DIA: JUEVES Y VIERNES **HORARIO: 19:00 HASTA 20:20**

CONTENIDO O TEMA A DESARROLLAR
Representación de los datos mediante tablas de frecuencias. Tablas de contingencia. Frecuencias simples, relativas y acumuladas. Representación de los datos mediante gráficos de barras, columnas, circulares e histogramas. Análisis de datos: medidas de tendencia central: media, moda y mediana. Cuidados y precauciones cuando se está en presencia de datos perdidos, datos anómalos y/o datos extremos.
GUIA O ACTIVIDADES
Lectura n. º 5: Organización de datos. Lectura n. ° 6: Representación de datos. Lectura n. º 7: Medidas de tendencia central. Trabajo Práctico n. º 3. Debe entregarse por medio de la opción “Realizar actividad” del Campus Virtual.
BIBLIOGRAFÍA
Está indicada al final de cada ficha de cátedra.
CANALES DE COMUNICACIÓN
Dudas y consultas a través de mensajería interna dentro del Campus Virtual (Educativa) de la Institución. Grupo de trabajo en Telegram: https://t.me/joinchat/UDvFVxxs7O2q2QHAHUAoAA Casilla de correo electrónico: iesestadistica@gmail.com

Se adjunta al presente material de estudio para el estudiante (de ser necesario).



FIRMA DEL DOCENTE

Lectura N° 5: Organización de datos

Palabras iniciales

Empezamos cuestiones prácticas. Recuerden que es imprescindible la comprensión de los conceptos teóricos previos. Sino le resultará demasiado incomprensible los ejercicios.

Organización de datos

Luego de aplicar algún instrumento de recolección de datos, ya sea entrevista, encuesta, cuestionario, tenemos un montón de datos que debemos organizar. Esta tarea se hace tediosa, cuanto más son las unidades de análisis, requiere mayor tiempo de organización, ni que hablar cuando los instrumentos son muy extensos. Por suerte, hoy en día hay muchas herramientas digitales que facilitan el proceso como Formularios Google, Survey Monkey, QuestionPro, etc.

Vamos a ver de manera práctica los conceptos cadena de datos, lote de datos y matriz de datos. Luego, 2 formas de organizar los datos, en tablas de frecuencias y tablas de contingencias.

Cadena de datos, lote de datos y matriz de datos

Vamos a analizar la encuesta que respondieron para poder darles el alta en la plataforma Educativa. Entre corchetes clasificaré según lo visto en la lectura n. º 4, en función del tipo de respuesta.

Continuamos aprendiendo

Correo electrónico: [abierta]

DNI/DU: [abierta]

Apellido: [abierta]

Nombre: [abierta]

Materia en la que se inscribe: [cerrada]

- Estadística (Gestión de Proyectos)
- Estadística (Comercialización)
- Estadística Aplicada (Recursos Humanos)

Condición: [cerrada]

- Me inscribí en la materia y curso por primera vez.
- Me inscribí en la materia y estoy recursando.
- No me inscribí en la materia porque estoy regular. Necesito refrescar conocimientos.

- No me inscribí en la materia porque no cumplo las correlativas. Quiero ser oyente.

Nota: la encuesta no se aplicó a estudiantes de Probabilidad y Estadística porque la institución proporcionó 1 semana antes del inicio de clases el listado de estudiante con email y número telefónico. En las otras instituciones sólo proporcionaron el DNI/DU de los estudiantes inscriptos, casi 2 semanas luego de iniciada las clases.

Como se confeccionó usando Formularios Google, se tiene una sistematización automática

Orden	Email	DNI/DU	Apellido	Nombre	Materia	Condición
1	estudiomucho1@correo.com	11111111	Einstein	Alberto	Estadística (GP)	Me inscribí en la materia y curso por primera vez.
2	estudiopoco1@correo.com	22222222	Giordano	Bruno	Estadística (C)	Me inscribí en la materia y estoy recursando.
3	noentiendonada@correo.com	33333333	Kant	Emanuel	Estadística Aplicada (RRHH)	Me inscribí en la materia y curso por primera vez.
4	seguroapruebo@correo.com	44444444	De Arco	Juana	Probabilidad y Estadística	Me inscribí en la materia y curso por primera vez.
	…	…	…	…	…	…
140	…	…	…	…	…	…

Si tuviéramos que reconocer varios conceptos que ya abordamos diremos que se trata de una investigación observacional, hay 6 variables donde todas son cualitativas. [El DNI/DU hace a la identificación única de las personas no es una expresión de cantidad o magnitud, es una suerte de "cualidad numérica"], el elemento de análisis es "*interesado* en cursar Estadística y afines" [porque hasta ese momento no tenía la confirmación de que fueran estudiantes, además dado que hay oyentes no debo limitar a estudiante], el criterio de variación es de interesado a interesado en cursar. Así podríamos seguir identificando conceptos, pero vamos a lo que interesa en esta lectura.

La cadena de datos es el conjunto de respuestas (datos) que otorga un elemento. Entonces, si pensamos en Emanuel Kant.

Cadena de datos 1 = {noenteindonada@correo.com; 33333333; Kant; Emanuel; Estadística Aplicada (RRHH); Me inscribí en la materia y curso por primera vez}

Podemos decir que la longitud de la cadena es 6, porque contiene 6 datos.

El lote de datos es el conjunto de respuestas (datos) que es agrupado por variable. Entonces, si pensamos en la variable "nombre del interesado en cursar Estadística y afines".

Lote de datos 1 = (Alberto; Bruno; Emanuel; Juana…) [Los puntos suspensivos indican que hay más datos. Son 140 nombres en total]

Podemos decir que el tamaño del lote de datos es 140. Como ya vimos en la lectura n. º 1, debo indicar si el tamaño pertenece a una muestra o población. Dado que los 140 no comprenden la totalidad se dice que $n = 140$.

La matriz de datos es el conjunto de respuestas (datos) que se dispone en un arreglo rectangular, donde cada columna es una variable. La matriz de datos representa al instrumento de recolección de datos. El cuadro que presenté arriba es "casi" una matriz de datos. Debería eliminar la columna "orden" porque no es una variable y ya sería una matriz de datos.

Pareciera ser un tema sencillo, pero como vimos en la lectura n. º 4, puede ocurrir muchas dificultades al momento de recolectar. Si es una encuesta escrita, puede que el encuestado no respondan algunas cuestiones, deje espacios en blanco o no indiquen el dato solicitado. Además, hay encuestas que suelen ser más.

Veamos un ejemplo más, en base a una encuesta más compleja.

Uso de Internet en hogares en niños en edad escolar

Para ser respondidas por tutores o padres

01. En relación con el niño/a, usted es:
    - Padre o madre biológica.
    - Padre o madre adoptiva.
    - Padre o madre de acogida. [Aclaración: son los que inician el proceso de adopción, ya conviven con los niños, pero aún no tienen sentencia firme de adopción].
    - Tutor/a legal.
02. Su hogar es:
    - Monoparental.
    - Biparental.
    - Otro tipo.
03. Dispone de conexión a Internet en su hogar.
    - Sí, permanentemente.
    - Sí, de manera esporádica.
    - No. [Fin de la encuesta]
04. ¿Qué le preocupa más sobre el uso de Internet respecto de sus hijos/as?
    - Que lo contacten extraños.
    - Que sufran un delito.
    - Que sean maltratados por otros niños.

- Que accedan a material inapropiado por Internet.
- Otras cosas… ¿cuáles?

05. Habitualmente, ¿desde dónde se conecta su hijo/a?
    - PC.
    - Teléfono móvil o Tablet.
    - Otros.
06. Su hijo/a dispone de una tarjeta espacial para compras por Internet con una cantidad de dinero mensual fija.
    - Sí.
    - No.

[Continua hasta la pregunta 17]

Noten que la pregunta 03, cumple la función de filtro. Dado que el objetivo es conocer sobre el uso de Internet en niños, si no disponen de dicha conexión no tiene sentido las siguientes preguntas. Por tanto, en 03, puede darse por finalizada la encuesta.

Si un niño no dispone de Internet en su hogar, la cadena de datos será

Cadena de datos = {Padre o madre biológica; Biparental; No}

La longitud es de 3.

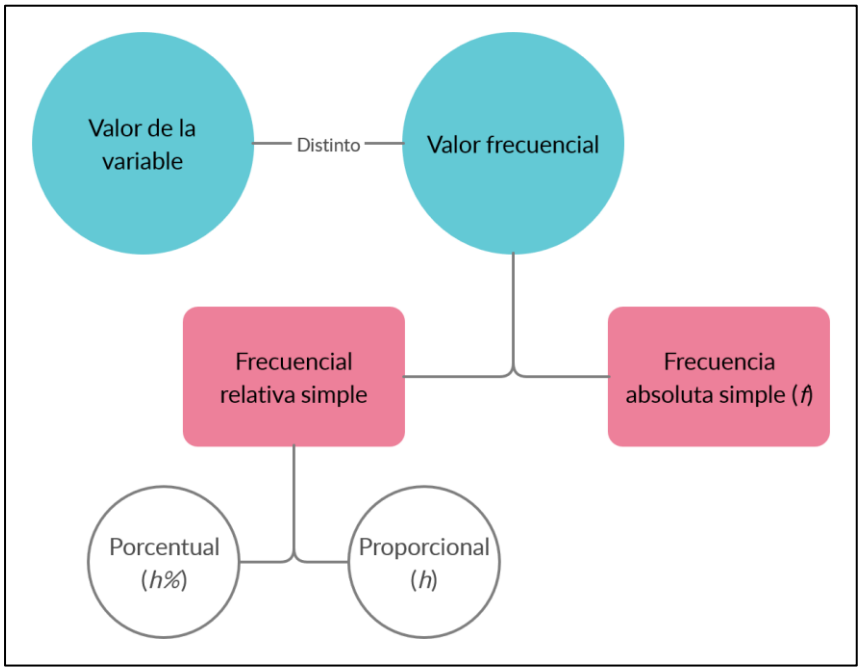
Lo último, muchas veces hay encuestas donde una respuesta conduce a otra pregunta y se van creando "huecos" en la matriz de datos. Esos huecos reciben el nombre de "datos vacíos". En dichos espacios no debería ir ningún dato. Sería ilógico pensar que luego de responder que no tiene conexión a Internet se le pregunte, "Habitualmente, ¿desde dónde se conecta su hijo/a?".

Tablas de frecuencias

Una tabla de frecuencia en una forma de organizar los datos de **una variable** en 2 componentes. Un componente que hace a los valores de la variable y un segundo componente formado por las frecuencias. La tabla de frecuencia enlista los valores de la variable cualitativa o de la variable cuantitativa discreta [con pocos valores], no así de variables cuantitativas continuas. ¿Por qué no continua?

Primer componente	Segundo componente		
Variable	*f*	*h*	*h%*
Valor 1	3	0,2120	21,20%
Valor 2	6	0,4240	42,40%
...	...	...	...
Total	n/N	1	100%

Éste es el error más común que muchos cometen en los primeros prácticos donde se pide reconocer la variable y sus valores. Los textos con los que trabajamos cuantifican los valores de la variable mediante proporciones o porcentajes, estos números les suele causar confusión.



Las frecuencias pueden ser simples o acumuladas. Por ahora nos dedicaremos a las frecuencias simples, las frecuencias acumuladas las veremos más adelante. La frecuencia absoluta simple o solo frecuencia (f) indica la cantidad de veces que se repite un dato. Es decir, tengo que contar cuantas veces se repitió, ese número irá en el valor de la variable correspondiente. La frecuencia relativa simple puede ser proporcional (h) o porcentual ($h\%$), es simplemente otra manera de expresar la cantidad de veces que se repitió un valor de la variable, pero con una transformación algebraica en función del tamaño de la muestra o población.

$$h = \frac{frecuencia\ absoluta}{tamaño\ del\ lote\ de\ dato}$$

$$h\% = \frac{frecuencia\ absoluta}{tamaño\ del\ lote\ de\ dato} * 100$$

Una cuestión más, cada fila de la tabla de frecuencias se llama "clase". En vez de decir, la fila 1, podemos decir la clase 1.

Por último, cada tabla que elaboren SIEMPRE debe llevar un título e indicarse la fuente de los datos. El título deberá estar contextualizado. Cuando se trabaja con fuentes de datos primarias se suele colocar "Fuente: elaboración propia", si se trabaja con fuentes de datos secundaria, se indica el nombre de la persona y su pertenencia organizacional, institución u organismo que posee los datos, por ejemplo "Fuente: INDEC", "Fuente: OMS" y "Fuente: Dra. María Salazar – CONICET".

Ejemplo de construcción de una tabla de frecuencia

Luego de aplicar la encuesta "Uso de Internet en hogares en niños en edad escolar" en 1006 adultos responsables al cuidado de niños en edad escolar. Como la encuesta capta 17 variables, se tendrían 17 tablas de frecuencias. Tomaremos la variable "Tipo de hogar constituido".

Lote de datos = (biparental; biparental; monoparental; monoparental; monoparental; monoparental; monoparental; biparental; biparental; biparental; biparental; biparental; biparental; biparental; biparental; biparental; biparental; biparental; biparental; monoparental; monoparental; monoparental; monoparental; monoparental; monoparental; biparental; biparental; otro; …) En teoría aquí se tiene 1006 datos.

Armemos la estructura. Una columna con el nombre de la variable y otras de frecuencias. No olvidar el título y la fuente. Hay que contar cuantas veces se repite un dato y colocarlo en la columna de frecuencia absoluta.

Tipo de hogar constituido en niños de edad escolar de España en junio 2014

Tipo de hogar	f	$h\%$
Biparental	707	70,3%
Monoparental	290	28,8%
Otro	9	0,9%
Total	1006	100%

Fuente: de elaboración propia

Nota: otro error común que cometen es nombra mal las fuentes. Si bien está encuesta la aplicó el Ministerio de Interior de España. Ellos recolectaron datos de primera mano. Por tanto, es una fuente primaria. Distinto sería que se tome esos datos para otra investigación. En ese caso seria fuente secundaria consignando "Fuente: Ministerio del Interior de España".

Tablas de frecuencias para variables cuantitativas

Cuando una variable cuantitativa discreta posea muchos valores, la construcción de la tabla de frecuencia sería muy extensa y poco práctica para analizar e interpretar. Cuando se trabaja con una variable cuantitativa continua no se podría construir porque sus valores son infinitos. Para estas situaciones, armaremos la tabla de frecuencia mediante intervalos.

Vamos a mostrar una tabla de frecuencias genérica de manera simbólica. Para poder redactar mejor y nombrar de manera inequívoca a cada frecuencia se acompaña con un subíndice numérico según la clase que ocupa.

Título: variable (+elemento) + espacio + tiempo

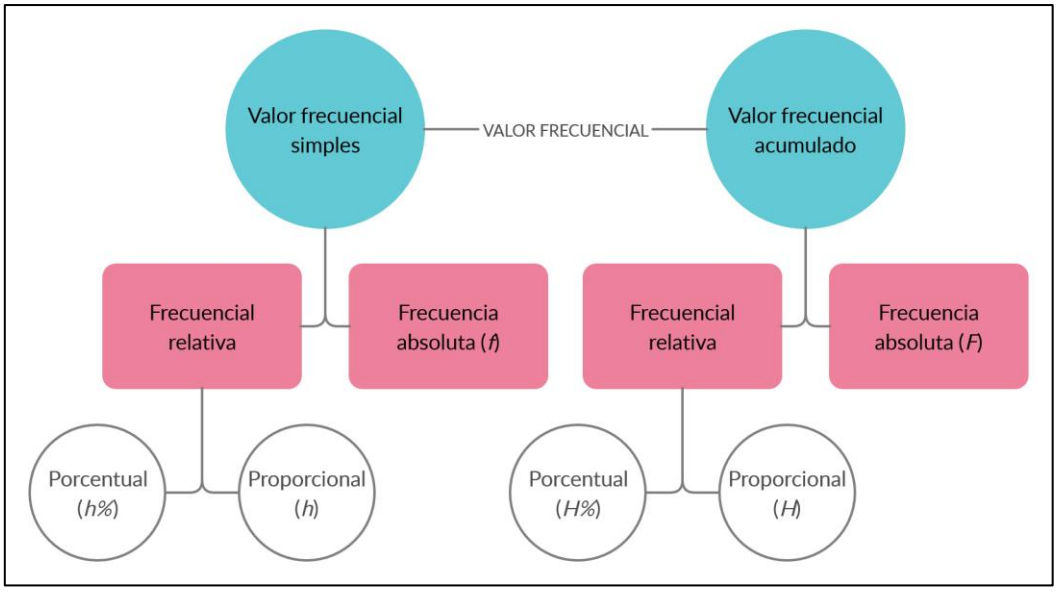
Variable	f	h	$h\%$	F	H	$H\%$
Valor/Intervalo	f_1	h_1	$h_1\%$	F_1	H_1	$H_1\%$
Valor/Intervalo	f_2	h_2	$h_2\%$	F_2	H_2	$H_2\%$
…	…	…	…	…	…	…
…	…	…	…	n/N	1	100%
Total	n/N	1	100%			

Fuente: primaria o secundaria.

Para obtener el número de intervalo y el ancho de cada intervalo haremos estos 3 pasos:

1. Aplicar la raíz cuadrada al tamaño del lote de datos o aplicar la regla de Sturges. El resultado lo llamaremos k, deben redondearlo. Este número me indica la cantidad de clases [filas] que debe tener la tabla de frecuencia.
   - $k = \sqrt{n}$ o $k = \sqrt{N}$, depende si trabajan con muestra o población.
   - $k = 1 + 3{,}322 * \log(n)$ o $k = 1 + 3{,}322 * \log(N)$, depende si trabajan con muestra o población.
2. Calcular el rango $[R]$ que es la diferencias entre el dato más grande, menos el dato más pequeño. $R = x_{máx} - x_{mín}$
3. Calcular el ancho del intervalo a. Siempre deben tomar el entero redondeado "hacia arriba" o que resulte más cómodo para trabajar. $a = R/k$

Hay que recordar que, en cada intervalo, no se toma el valor superior, excepto en la última clase, donde el último intervalo es cerrado. Cuando trabajamos con variables cuantitativas podemos trabajar con las frecuencias acumuladas. La simbología es la misma que las frecuencias simples, pero en mayúsculas.



¿Qué quiere decir que sean acumuladas? Que van a ir sumando consecutivamente los valores de la variable. Por eso las frecuencias acumuladas carecen de sentido práctico cuando se trabaja con variables cualitativas (juntar dos cualidades muchas veces carece de sentido). Entonces F, H y $H\%$ se obtendrá sumando f, h y $h\%$ respectivamente. Esa suma es acumulativa con la frecuencia inmediata anterior. A fines didácticos se mostrará cómo se va sumando la columna F, en una tabla de frecuencia sólo deben colocar los resultados. Igual procedimiento se aplica en H y $H\%$.

Edad de vecinos del barrio X. Ciudad de Salta, octubre 2020

Intervalos de edades	f	h	$h\%$	F	H	$H\%$
[15 - 25)	11	0,3667	36,67	11	0,3667	36,67%
[25 - 35)	10	0,3333	33,33	21 [11+10]	0,7000	70,00%
[35 - 45)	4	0,1333	13,33	25 [21+4]	0,8333	83,33%
[45 - 55)	2	0,0667	6,67	27 [25+2]	0,9000	90,00%
[55 - 65]	3	0,1000	10,00	30 [27+3]	1	100%
Total	30	1	100			

Las precauciones que debe tener en la construcción es asegurarse que la columna de frecuencias acumuladas no se totaliza, por eso rellené con negro la última fila de dichas columnas. La penúltima fila de frecuencias absoluta acumulada debe ser n/N, de la frecuencia relativa proporcional acumulada 1 y de la frecuencia relativa porcentual acumulada 100%.

Interpretemos algunas frecuencias acumuladas, por ejemplo, $F_4 = 27$, quiere decir que 27 personas tienen entre 15 y 55 años. $H_2\% = 70\%$, quiere decir que el 70% de las 30 personas tienen entre 15 y 35 años. Cuando se debe interpretar una frecuencia acumulada, deben extender o ampliar el intervalo.

Tablas de contingencia

Una tabla de contingencia lo que hace es "fusionar" 2 tablas de frecuencias, permitiendo visualizar si existe algún tipo de relación entre ambas variables que llame la atención al investigador.

El único cuidado que deben tener en su construcción es que la tabla esté balanceada. ¿Qué quiere decir balanceado? Que el lote de la variable 1, sea igual al lote de la variable 2. $n_1 = n_2$.

Como están combinando/fusionando 2 variables, deben procurar ese balance. Si no tienen este cuidado, tendrán tablas desbalanceadas, que no aporta información para la toma de decisiones porque algunas sumas serán ilógicas.

En su diseño, se cruzan los valores de ambas variables obteniendo frecuencias dobles. El margen horizontal o total por fila es la suma de las filas. La suma de las columnas da por resultado el margen vertical o total por columna. La suma de los márgenes da por resultado el tamaño del lote de datos [n o N].

Importante: a diferencias de la tabla de frecuencia que puede trabajar con una sola frecuencia o con varias frecuencias simultáneamente, cada tabla de contingencia trabaja con un solo tipo de frecuencia.

Para identificar cada frecuencia en la tabla de contingencia se le agrega 2 subíndices numéricos, el primero indica la fila y el segundo la columna. Por ejemplo, f_{32} indica la frecuencia de la fila 3 y de la columna 2.

Título: variable 1 [nexo] variable 2 (+ elemento) + tiempo + lugar

		Variable 2			Margen vertical o total por columna
		Valor 1	Valor 2	Valor …	
Variable 1	Valor 1	f_{11}	f_{12}	…	f_1
	Valor 2	f_{21}	f_{22}	…	f_2
	Valor …	…	…	…	…
Margen horizontal o total por fila		f_1	f_2	…	n/N

Fuente: primaria o secundaria

Ejemplo de aplicación:

Al finalizar el examen final del segundo llamado de julio 2020, el docente consulta a los 20 estudiantes que rindieron, sobre su intensidad de estudio en función de un rango de horas, previo para dicho examen.

Luego, de organizar las respuestas se muestra la siguiente tabla:

Variable	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
Resultado	Aprobó	Aprobó	Reprobó	Reprobó	Aprobó	Aprobó	Aprobó	Reprobó	Reprobó	Reprobó
Intensidad de estudio	Mucho	Más o menos	Poco	Más o menos	Más o menos	Mucho	Mucho	Mucho	Más o menos	Poco

Variable	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20
Resultado	Reprobó	Aprobó	Reprobó	Aprobó	Reprobó	Aprobó	Aprobó	Reprobó	Aprobó	Reprobó
Intensidad de estudio	Mucho	Mucho	Poco	Mucho	Más o menos	Poco	Mucho	Más o menos	Mucho	Poco

Al cruzar las 2 variables dan lugar a 6 categorías, aprobó estudiando poco, aprobó estudiando más o menos, aprobó estudiando mucho, reprobó estudiando poco, reprobó estudiando más o menos y reprobó estudiando mucho. Observo la tabla y voy contando en función de las categorías. Así, por ejemplo, C3, C10, C13 y C20 son todos los casos para la categoría reprobó estudiando poco.

Relación entre los resultados del examen final y la intensidad de estudio por parte de los estudiantes del Instituto … en el segundo llamado de julio de 2020.

Intensidad de estudio / Resultados	Poco	Más o menos	Mucho	Margen vertical
Aprobó	1	2	7	10
Reprobó	4	4	2	10
Margen horizontal	5	6	9	20

Fuente: elaboración propia.

También pueden transformar esta tabla de contingencia con frecuencias absolutas en frecuencias relativas, de manera similar que en las tablas de frecuencias.

Intensidad de estudio / Resultados	Poco	Más o menos	Mucho	Margen vertical
Aprobó	5%	10%	35%	50%
Reprobó	20%	20%	10%	50%
Margen horizontal	25%	30%	45%	100%

Fuente: elaboración propia.

¿Qué les llama la atención de la tabla de contingencia? Siempre es bueno observar lo más frecuente y lo menos frecuente. Así, por ejemplo, los que aprobaron son los que estudiaron mucho y es raro o poco frecuente, aprobar estudiando poco.

Las tablas de contingencia ayudan a comprender si es que existe algún tipo de relación entre las variables. Son muy útiles en una investigación.

Palabras de cierre

Cómo habrán visto, cuando aplicamos un instrumento de recolección se realizan muchas preguntas (u observaciones) y cada pregunta está relacionada con una variable. Es interesante que el investigador "juegue" a relacionar variables para ver si halla algo interesante.

Bibliografía

Gorgas García, J.; Cardiel López, N. y Zamorano Calvo, J. (2011). Estadística básica para estudiantes de ciencias. Editorial de la Universidad Complutense de Madrid, España.

Lectura N° 6: Representación de datos

Palabras iniciales

Vamos a continuar con las cuestiones prácticas. Recuerden que es imprescindible la comprensión de los conceptos teóricos previos, sino le resultará demasiado incomprensible los ejercicios. Lo que vamos a hacer es transformar las tablas en gráficos.

Gráficos

En esta lectura abordaremos el tema de gráficos. Para hacer los gráficos, debemos tener y contar con los datos organizados. Por ello, debemos construir primero las distintas tablas que necesitemos acorde con la investigación y en función de ellas hacer el gráfico. Permiten la visualización de los datos.

Vamos a organizar el tema de la siguiente manera:

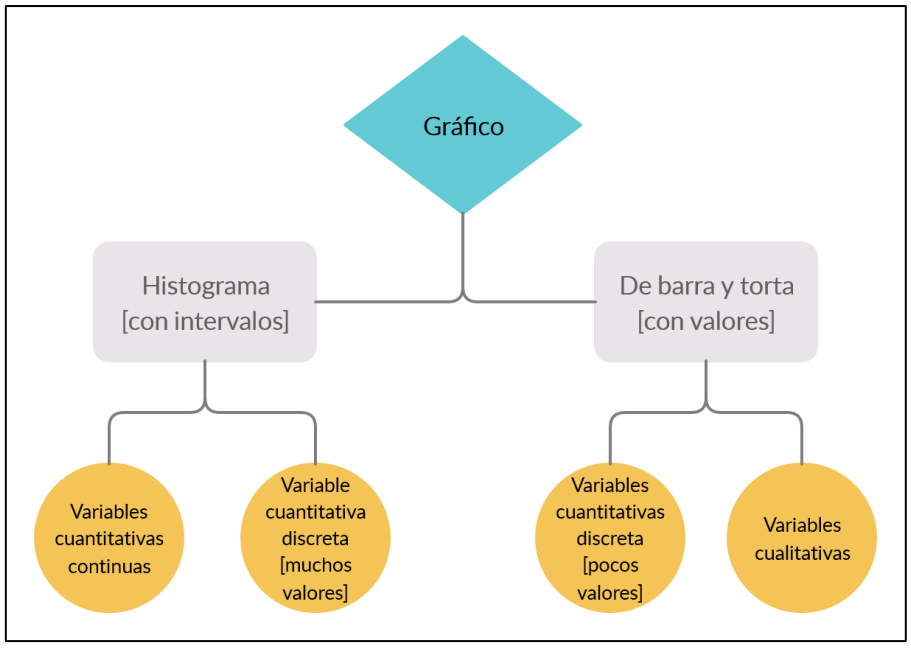
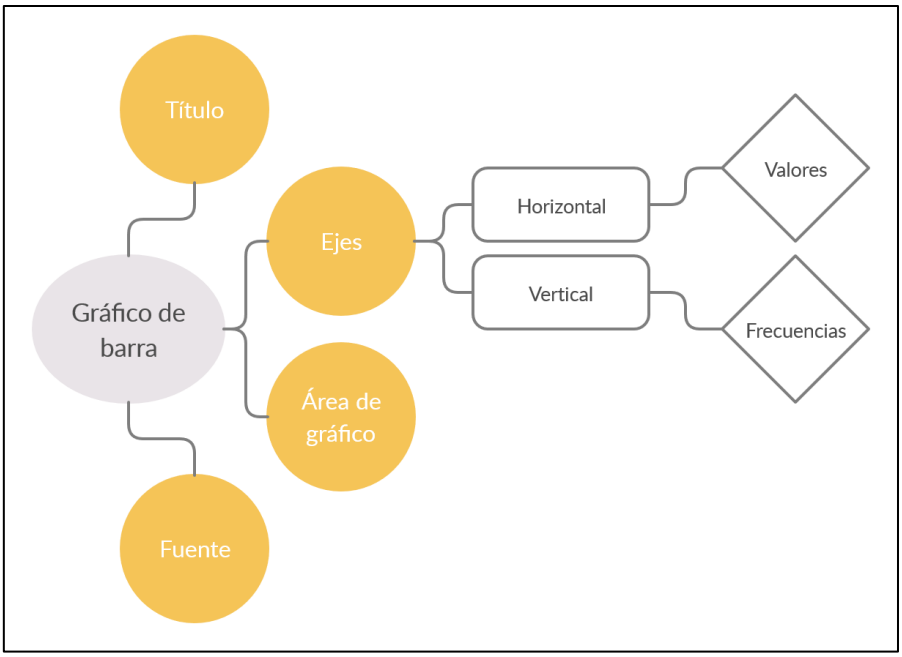


Gráfico de barras

El gráfico de barras o columnas sirve para representar visualmente la frecuencia con la que las variables cualitativas y variables cuantitativas discretas con pocos valores ocurren.

Los elementos infaltables serán el título, los ejes, el gráfico propiamente dicho y la fuente de datos.

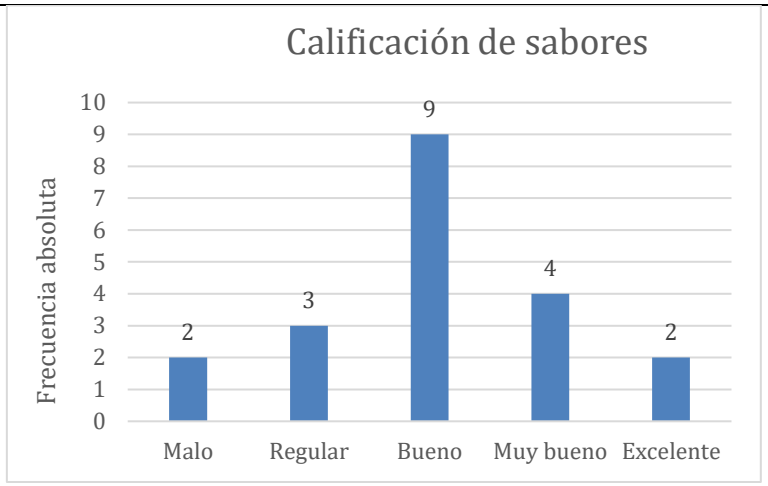


En el eje horizontal deberán ubicar los valores de las variables, mientras que en el eje vertical decidirán cuál de las frecuencias simples quieren usar.

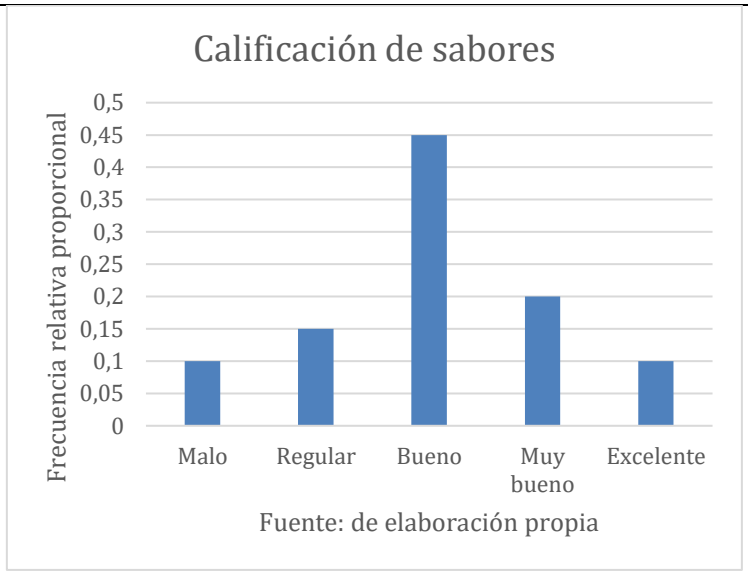
Ejemplo de gráfico de barra

La siguiente tabla la pasaremos a gráfico.

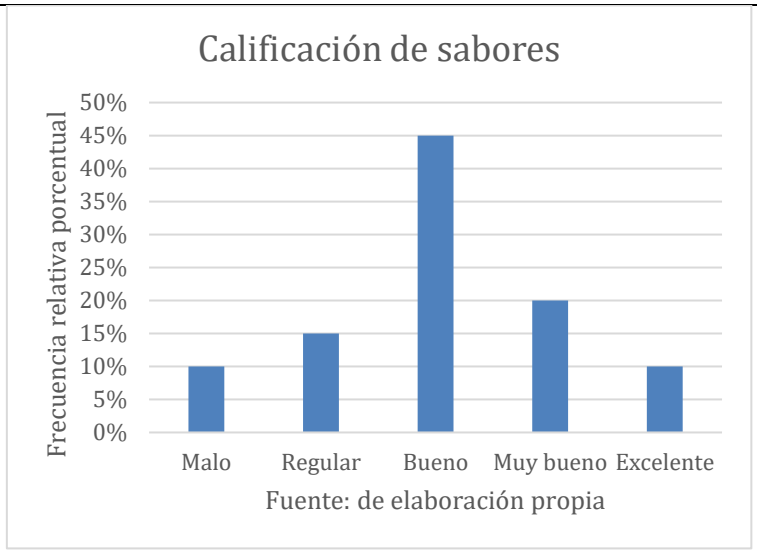
Calificación de sabores	f	h	$h\%$
Malo	2	0,1	10%
Regular	3	0,15	15%
Bueno	9	0,45	45%
Muy bueno	4	0,2	20%
Excelente	2	0,1	10%
Total	20	1	100%



Fuente: de elaboración propia



Se trabajó con una muestra de 20



Se trabajó con una muestra de 20

Noten que las 3 gráficas son similares, lo único que cambia es la escala del eje vertical que coindice con la frecuencia simple que elijan, puede ser f, h o $h\%$.

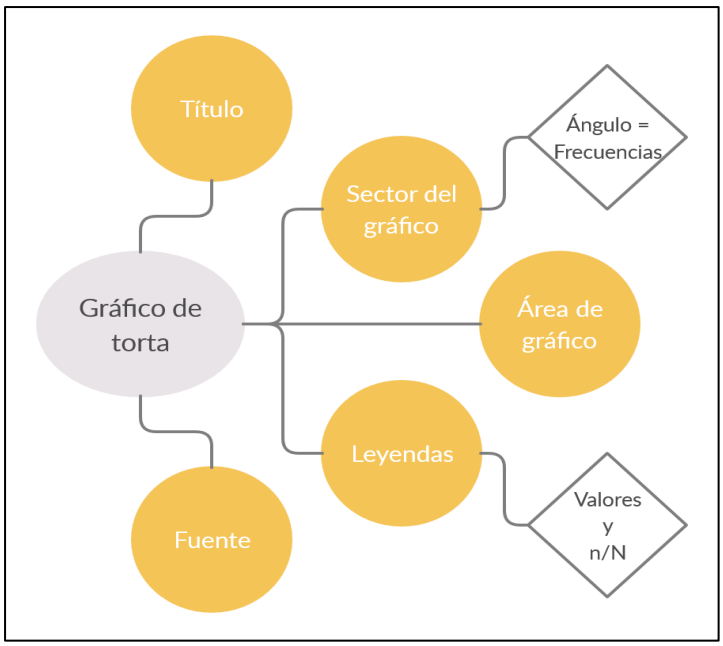
Gráfico de torta

El gráfico de torta o circular sirve para representar visualmente la frecuencia con la que las variables cualitativas y variables cuantitativas discretas con pocos valores ocurren.

Los elementos infaltables serán el título, los sectores, el gráfico propiamente dicho, la leyenda y la fuente de datos.

Para hacer el gráfico, debemos transformar las frecuencias en ángulos. Para ello usaremos la regla de 3 simple

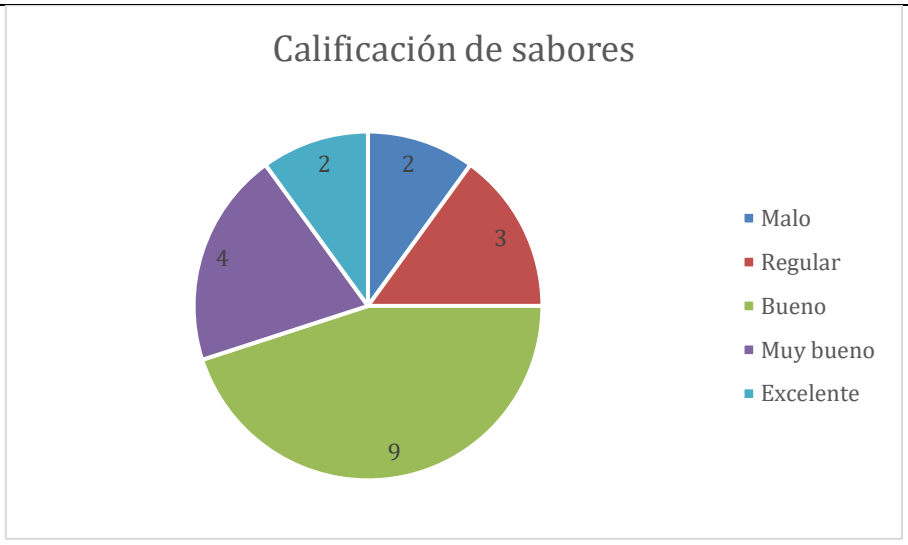
$$n/N \quad \_\_\_\_\_ \quad 360°$$
$$f/h/h\% \quad \_\_\_\_\_ \quad x°$$



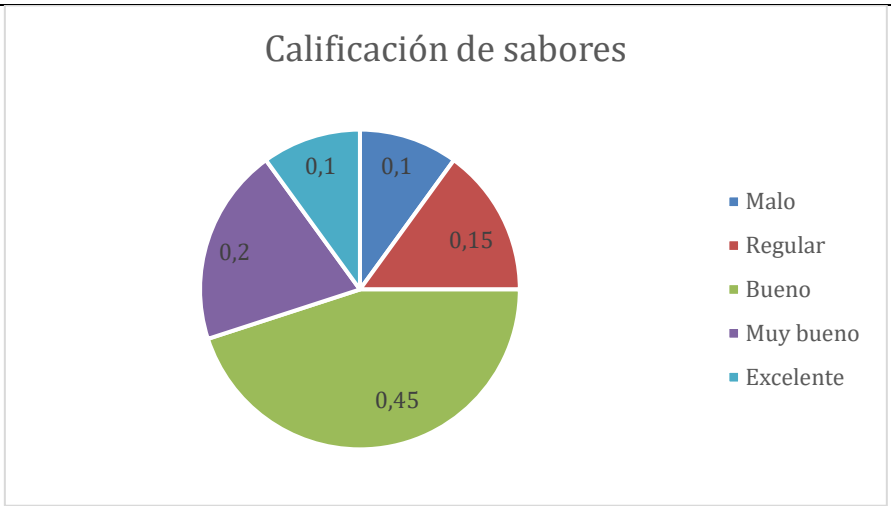
Ejemplo de gráfico de torta

La siguiente tabla la pasaremos a gráfico.

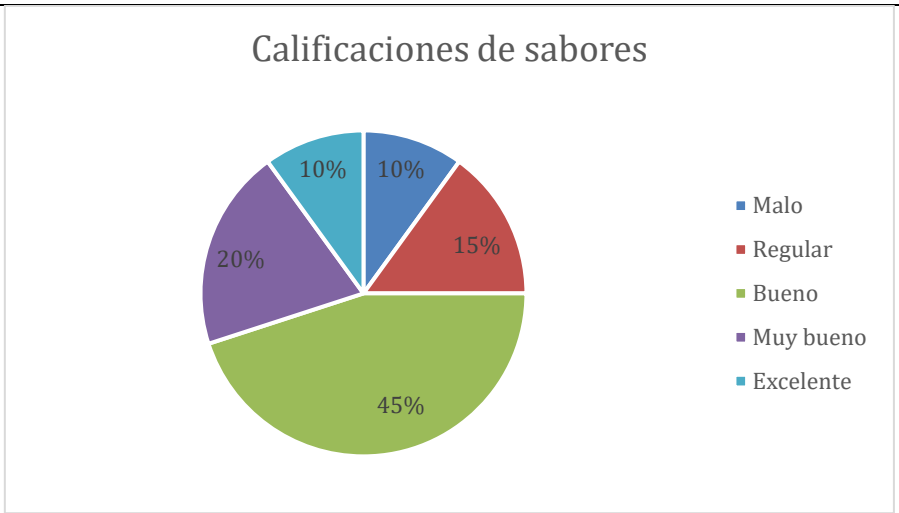
Calificación de sabores	f	h	$h\%$
Malo	2	0,1	10%
Regular	3	0,15	15%
Bueno	9	0,45	45%
Muy bueno	4	0,2	20%
Excelente	2	0,1	10%
Total	20	1	100%



Fuente: elaboración propia



Fuente: elaboración propia. Se trabajó con una muestra de 20.



Fuente: elaboración propia. Se trabajó con una muestra de 20.

Por ejemplo, el ángulo del sector naranja será 54°. Recuerden que podemos trabajar con regla de 3 simple.

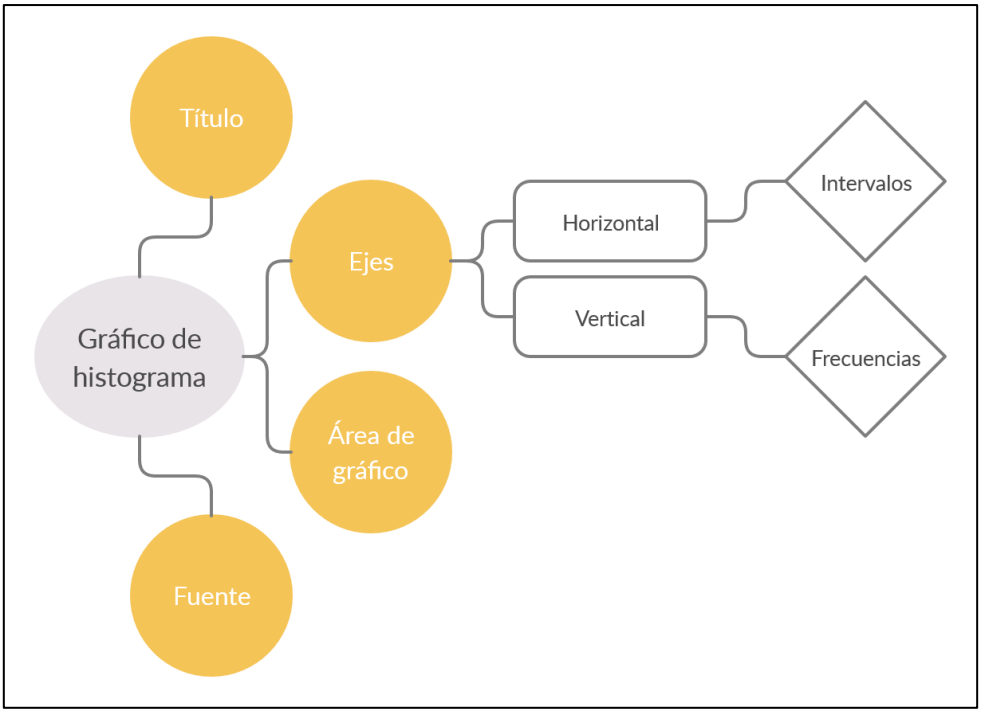
$$20 \_\_\_\_ 360°$$

$$3 \_\_\_ x°$$

$$x = \frac{3 * 360°}{20} = 54°$$

Histograma

El histograma es muy similar a un gráfico de barras, pero al tratarse de un caso para representar intervalos, el eje horizontal será numérico.



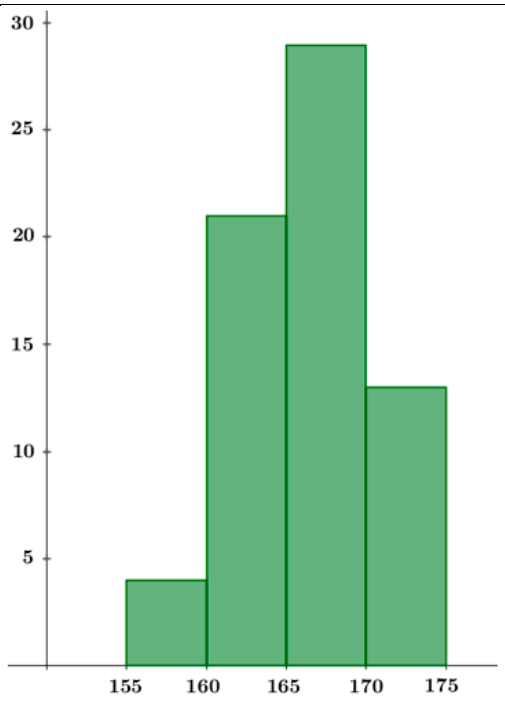
Ejemplo de histograma

Vamos a representar la siguiente tabla de frecuencia.

Estatura	f_i
[155, 160)	4
[160, 165)	21
[165, 170)	29
[170, 175)	13

Recuerden que, al trabajar con intervalos, pierdo la individualidad de los datos. No sé cuántos estudiantes tienen una altura de 1,63, pero sí sé que son 21 los que están contenidos en dicho intervalo.

Estatura de los estudiantes de primer año de una universidad pública. Marzo 2019.



Fuente: elaboración propia. Se trabajó con una muestra de 67.

Apéndice sobre intervalos

Representación simbológica de intervalos.

Intervalo cerrado: se usa corchetes [ ].

Ejemplo: $[1;\ 3{,}5]$ quiere decir que contiene todos los números desde 1 a 3,5 y los extremos pertenecen al conjunto.

Intervalo abierto: se usa paréntesis ().

Ejemplo: $(1;\ 3{,}5)$ quiere decir que contiene todos los números desde 1 a 3,5 pero los extremos no pertenecen al conjunto.

Intervalo semiabierto o semicerrado: [ ) o ( ]

Ejemplo: $[1;\ 3{,}5)$ quiere decir que contiene todos los números desde 1 a 3,5. El 1 pertenece al conjunto, pero 3,5, no pertenece.

Ejemplo: $(1;\ 3{,}5]$ quiere decir que contiene todos los números desde 1 a 3,5. El 1 no pertenece al conjunto, pero 3,5, sí pertenece.

Recuerden que los histogramas y las tablas de frecuencias con intervalos lo construimos con intervalos semiabiertos [ ), salvo la última clase que puede ser de intervalo cerrado [ ]. Dependerá del lote de datos en particular, el ancho del intervalo y de verificar que todos los datos estén contenidos en él.

Retomando la tabla de frecuencia anterior, si tengo el dato (1,60) pertenecerá a la segunda clase, porque la primera clase no lo contiene por ser abierta por derecha. Si tengo el dato (1,70), pertenecerá a la cuarta clase. En este caso particular, nadie media 1,75, por ello no fue necesario que se cierre el último intervalo.

Estatura	f_i
[155, 160)	4
[160, 165)	21
[165, 170)	29
[170, 175)	13

.

Palabras de cierre

Hasta aquí terminamos las cuestiones que hacen a la organización y representación de los datos. Las próximas 2 lecturas abordaran el tema de análisis de datos para variables cuantitativas exclusivamente.

Bibliografía

Gorgas García, J.; Cardiel López, N. y Zamorano Calvo, J. (2011). Estadística básica para estudiantes de ciencias. Editorial de la Universidad Complutense de Madrid, España.

Lectura N° 7: Medidas de tendencia central

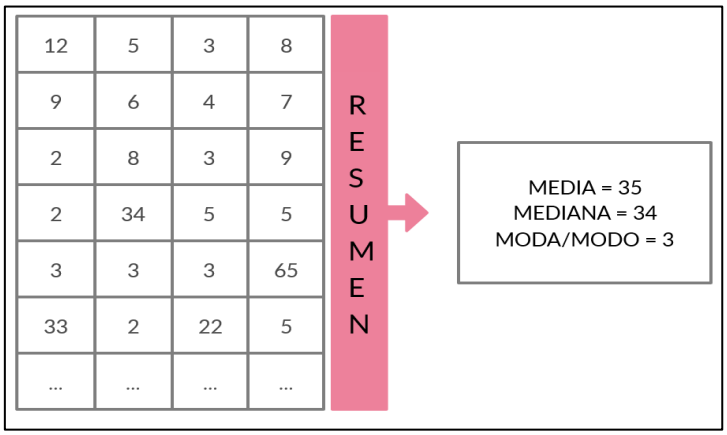
Palabras iniciales

En esta lectura y próxima vamos a centrarnos exclusivamente en técnicas que sólo sirven para variables cuantitativas. Dado que se tratan de cálculos, nos centraremos en la interpretación, ya que los softwares y calculadoras harán esto por nosotros.

Medidas de tendencia central

Hay muchas medidas de tendencia central. Nosotros nos centraremos en 3 medidas básicas: la media, la mediana y el modo o moda. Estos son números resúmenes de un lote de datos. Lo que vamos a hacer es reducir un gran lote de datos a estos números [estadísticas]. Es lo que en el video de bienvenida de Walter Escudero expresa como simplificaciones útiles. Perdemos la individualidad de los datos y empezamos a hablar

en términos del conjunto. A modo de ejemplo, presentamos el siguiente esquema, estos números de tendencia central son representantes de todo el lote de datos.



Simbolización de las medidas de tendencia central

Vamos a presentar más simbología a la que ya venimos aprendiendo. Aquí vamos a diferenciar los símbolos según se trate de estadísticos o parámetros.

Medida	Estadístico	Parámetro
Media	$\bar{x}$ [se lee equis barra]	μ [letra griega "mu"]
Mediana	$\tilde{x}$ [se lee equis lomita] o Me	No hay acuerdo simbólico, podemos usar $\tilde{X}$ (mayúscula).
Modo/moda	$\dot{x}$ [se lee equis punto] o Mo	No hay acuerdo simbólico, podemos usar $\dot{X}$ (mayúscula).

Media [aritmética]

Usualmente se la confunde con la palabra "promedio". Un promedio es algo que más o menos está en el medio, pero necesitamos ser específicos, por lo cual la palabra correcta es media aritmética o simplemente media. En los diarios se usual usar promedio como sinónimo de media, pero no es adecuado.

La media se la define como la suma de los datos, dividido entre el tamaño del lote de datos. Geométricamente representa el punto de equilibrio de la suma de distancias entre cada dato y la media, separando entre aquellos datos superiores a la media y los datos inferiores a la media. El símbolo Σ [letra griega mayúscula "sigma"] simplifica una secuencia de sumas, podemos entenderla como "sumar todo".

$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$	$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$

Ejemplo 1: el siguiente lote de datos, representa las edades de una muestra de 12 ingresantes al terciario: $(21; 17; 18; 20; 21; 26; 45; 19; 18; 19; 22; 21)$. Ahora calculamos la media.

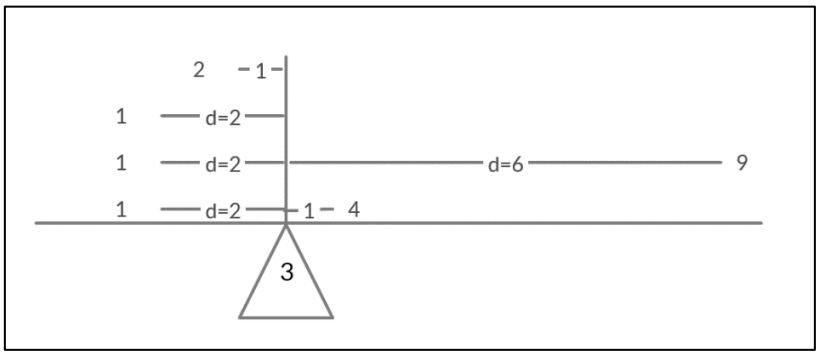
$$\bar{x} = \frac{21 + 17 + 18 + 20 + 21 + 26 + 45 + 19 + 18 + 19 + 22 + 21}{12} = \frac{267}{12} = 22{,}25$$

Podemos decir que la edad media de ese grupo de 12 personas es 22,25 [Pregunta de repaso, ¿qué tipo de estadística es 22,25?]. Recuerden, este número resume todo el lote de datos, perdiéndose la individualidad

de cada dato. Sería imposible escribir un informe, una noticia o un artículo si no hacemos estas reducciones necesarias, más aún cuando se trabajan con miles de datos.

Ejemplo 2: este lote representa la cantidad de hermanos de 6 alumnos. $(1; 2; 4; 9; 1; 1)$.

$$\bar{x} = \frac{1+2+4+9+1+1}{6} = \frac{18}{6} = 3$$



Noten que, si lo representamos geométricamente, el 3 representa el punto de equilibrio de distancias superiores e inferiores respecto a la media. Si sumamos las distancias de los datos 1, 1, 1, y 2, respecto a la media, suma una distancia de 7. Si sumamos las distancias de los datos 4 y 9, respecto a la media, suma una distancia de 7.

Mediana o valor medio

La mediana o valor medio es el dato que se ubica en el centro de un lote de datos ordenado. Tendremos 2 situaciones. Cuando el tamaño del lote de datos es par o impar.

- Impar

$$\tilde{x} = x_{(\frac{n+1}{2})}$$

- Par

$$\tilde{x} = \frac{x_{(\frac{n}{2})} + x_{(\frac{n+2}{2})}}{2}$$

Para calcular:

1. Tendremos que ordenar el lote de datos de manera ascendente.
2. Calcular según si el lote de datos es par o impar.
3. Reemplazar acorde con la ubicación del dato.

Ejemplo 1: el siguiente lote de datos, representa las edades de una muestra de 12 ingresantes al terciario: $(21; 17; 18; 20; 21; 26; 45; 19; 18; 19; 22; 21)$. Ahora calculamos la mediana.

Ordenamos de menor a mayor.

$$(17\ ;18\ ;18\ ;19\ ;19\ ;20\ ;\ 21\ ;21;21\ ;22\ ;\ 26\ ;\ 45)$$

$$x_1\ ;\ x_2\ ;\ x_3\ ;x_4\ ;\ x_5\ ;\ x_6;\ x_7\ ;\ x_8\ ;\ x_9;\ x_{10}\ ;\ x_{11}\ ;\ x_{12}$$

Como el lote es par, $n = 12$.

$$\tilde{x} = \frac{x_{(\frac{n}{2})} + x_{(\frac{n+2}{2})}}{2} \rightarrow \tilde{x} = \frac{x_{(\frac{12}{2})} + x_{(\frac{12+2}{2})}}{2} \rightarrow \tilde{x} = \frac{x_{(6)} + x_{(7)}}{2}$$

$$\tilde{x} = \frac{20 + 21}{2} \rightarrow \tilde{x} = \frac{41}{2} = 20{,}5$$

Concluimos que 20,5 es el valor medio del lote de datos. La cantidad de datos inferior a 20,5 es igual a la cantidad de datos superior a 20,5. Hay 6 datos inferior a la mediana y 6 datos superior a la mediana.

Ejemplo 2: este lote representa la cantidad de hermanos de 5 alumnos. $(1; 2; 4; 9; 1)$.

Ordenamos de menor a mayor.

$$(1\ ;\ 1\ ;\ 2;\ \ 4\ ;\ 9)$$

$$x_1; x_2; x_3; x_4; x_5$$

Como el lote es impar, $n = 5$.

$$\tilde{x} = x_{(\frac{n+1}{2})} \rightarrow \tilde{x} = x_{(\frac{5+1}{2})} \rightarrow \tilde{x} = x_{(3)}$$

$$\tilde{x} = 2$$

Tenemos que 2 es la mediana del lote de datos. Hay 2 datos inferior y hay 2 datos superior.

Si la media es el punto de equilibrio de las distancias, la mediana es el punto de equilibrio de cantidades.

Moda o modo

La moda es el mayor valor frecuencial de un conjunto de datos. La moda se utiliza de manera más común para datos cualitativos que para datos cuantitativos, cuando ni la media ni la mediana pueden usarse con datos cualitativos.

$$Mo = f_{máx} = h_{máx} = h\%_{máx}$$

Si **todos** los datos tienen el mismo valor frecuencial, entonces se concluirá que no hay moda o modo. Si el lote de datos tiene una sola moda, se dice que tiene distribución unimodal. También se puede dar donde más de un dato comparta el mayor valor frecuencia, en dicho caso se dirá que la moda es multimodal.

Ejemplo 1: la moda del lote de datos $\dot{x} = 21$, porque tiene el mayor valor frecuencial $f_{mo} = 3$

$$(17\ ;18\ ;18\ ;19\ ;19\ ;20\ ;\ 21\ ;21;21\ ;22\ ;\ 26\ ;\ 45)$$

No hay fórmula, surge de inspección y observación. Si tengo los datos representados en una tabla de frecuencia observo la mayor frecuencia.

No confundir el valor modal con el valor frecuencial. En el ejemplo 1, se suele cometer el error de $Mo = 3$, pero 3 no es un dato.

Ejemplo 2: la moda del siguiente lote es $Mo = Rojo$.

$$(Negro; amarillo; rojo; verde; verde; rojo; rojo; azul)$$

Ejemplo 3: no hay moda. Todos los datos tienen el mismo valor frecuencial.

$$(Negro; amarillo; rojo; verde; azul)$$

Ejemplo 4: la moda es $Mo = \{rojo; verde\}$

$$(Negro; rojo;\ amarillo; rojo; verde; azul; verde)$$

Datos atípicos y la representatividad de la media

Un dato atípico de un conjunto de datos es el dato que es mucho mayor o menor que la mayoría de los demás datos.

Por ejemplo: la media del lote 1 $(1; 2; 2; 3; 5)$ es $\bar{x} = 2{,}6$. La media del lote 2 $(1; 2; 2; 3; 45)$ es $\bar{x} = 10{,}6$. Noten que ni la mediana ni modo cambian porque los datos atípicos no las afectan.

Vemos que, al alterar el valor de un dato, la media sufre un cambio drástico. Entonces, la media es sensible a los datos atípicos. Esto nos lleva a pensar, ¿qué tan representativo es la media de un conjunto de datos?

Tal vez 2,6 más o menos representa a todo el lote 1, pero 10,6 no es un buen representante del lote 2. Para saber si la media es buena medida resumen, necesitamos calcular la medida de homogeneidad. Esto lo veremos en la siguiente lectura luego de introducir el tema de dispersión.

Recuerden que trabajar con una muestra no implica que pueda hacer una generalización sobre la población. La muestra debe ser representativa.

Resumen

	Media	Mediana	Moda
Definición	Equilibrio de distancias.	Dato central [lote de datos impar] o datos centrales [lote de datos par]	Dato o datos más frecuentes.
¿Qué tan común es en artículos o noticias?	Muy común. El más familiar.	Común.	Poco usado.
¿Existe en un lote de datos cuantitativos?	Siempre.	Siempre.	Podría no haber moda o más de una.
¿Existe en un lote de datos cualitativos?	Nunca.	Nunca.	Podría no haber moda o más de una.
¿Toma en cuenta todos los datos?	Sí.	No (sólo 1 o 2 centrales).	No.
¿Afectada por datos atípicos?	Sí.	No.	No.
Ventajas	Funciona bien con muchos métodos estadísticos.	Es mejor que la media cuando hay datos atípicos.	La mejor para comprender el comportamiento de datos cualitativos.

Palabras de cierre

Hasta aquí terminamos las medidas para poder resumir un lote de datos cuantitativo. Con la próxima lectura finalizaremos el análisis de datos. Luego iniciaremos algunas nociones de probabilidad.

Bibliografía

Gorgas García, J.; Cardiel López, N. y Zamorano Calvo, J. (2011). Estadística básica para estudiantes de ciencias. Editorial de la Universidad Complutense de Madrid, España.

Trabajo Práctico N. ° 3

Lea atentamente las consignas y responda luego de la lectura n. ° 4 y 5. Responda preferentemente a continuación de cada actividad.

Para las 2 encuestas que tienen a continuación armen un bosquejo de la matriz de datos. No es necesario que recolecten datos. Solamente deben presentar la estructura. Para un ejemplo, revise la lectura n. º 5.

RECORDAR: la variable se expresa en sentido afirmativo impersonal.

Encuesta n. º 1

01. ¿Por qué te pusieron tu nombre de pila?
02. Si tuvieras que promocionarte para un trabajo, ¿cómo sería tu slogan?
03. Si fueras director de cine, ¿a qué género pertenecería tu película?
04. Si retrocedieras al momento de elegir una carrera, ¿elegirías la misma?
05. ¿Qué animal mítico te gustaría ser?
06. Define con una palabra lo que significa Internet para ti.

Encuesta n. º 2

01. ¿En qué año nació?
02. ¿Cuál es el último nivel de formación académica que completó?
03. ¿Trabajó la semana pasada?
    - Sí. [Continuar en 05]
    - No. [Continuar en 04]
04. ¿No trabajó por alguno de los siguientes motivos?
    - Está de vacaciones. [Continuar en 05]
    - Está usufructuando alguna licencia. [Continuar en 05]
    - Está suspendido de su trabajo. [Continuar en 05]
    - Es jubilado. [Continuar en 05]
    - No tengo trabajo. [Continuar en 08]
05. ¿Cuántas horas de trabajo realiza semanalmente?
06. La franja horaria de trabajo es principalmente:
    - Matutino. [Continuar en 08]
    - Tarde. [Continuar en 08]
    - Vespertino. [Continuar en 08]

- Nocturno. [Continuar en 08]
- No es posible indicar una respuesta. [Continuar en 07]

07. La condición horaria de trabajo es:
    - Con más de una franja horaria.
    - Rotativa según determinados días.
    - Alternada por semanas.
    - Rotativa y alternada.
08. Dada las circunstancias actuales se pronostica un crecimiento de demanda laboral al sector de servicios y asesorías. Le gustaría iniciar una nueva carrera relacionada a la programación y ciencia de los datos.
    - Sí. [Continuar en 09]
    - No. [Fin de la encuesta]
09. ¿Qué horario de cursado le favorece?
    - Matutino.
    - Vespertino.

Clasifique cada pregunta de la encuesta n. º 2, bajo el criterio "contestación del encuestado" y "contenido de la pregunta". A modo de ejemplo, se clasifica la pregunta 01.

Pregunta 01: Abierta de identificación.

Pregunta 02:

Pregunta 03:

Pregunta 04:

Pregunta 05:

Pregunta 06:

Pregunta 07:

Pregunta 08:

Pregunta 09:

Le encomendaron revisar las respuestas de la encuesta nº 2, del inciso 1. Señale si las cadenas de datos se deben descartar o están bien. A modo de ejemplo, se realiza el inciso a. Los espacios vacíos indican que ahí no debería ir ningún dato, si les resulta práctico pueden escribir la palabra "vacío".

a. Cadena de datos 1 = {18 años; secundaria completa; no; no tengo trabajo; ; ; ; sí; matutino}.

El primer dato es incorrecto. Se pregunta por el año de nacimiento no la edad actual. No la descartaría simplemente calcular 2020-18 e indicar que nació en el 2002.

b. Cadena de datos 2 = {2000; secundaria completa; sí; ; 20 horas semanales; vespertino; ; no; }
c. Cadena de datos 3 = {1998; secundaria completa; no; ; 15 horas semanales; no es posible indicar una respuesta; rotativa y alternada; no; matutino}

Indique para la siguiente situación si la encuesta es adecuada o no. Señale los errores conceptuales si los hubiera.

Planeamiento del problema: está colaborando en el ingreso de estudiantes a la carrera de Soporte y Mantenimiento Informático, 2021. Para estimar la posible cantidad de ingresantes ante la incertidumbre del contexto actual, elije 15 colegios de gestión pública de la zona céntrica y oeste de la ciudad de Salta. Mediante los directivos de dichas instituciones les hace llegar la siguiente encuesta a los estudiantes para que respondan de manera online.

Encuesta
[01] Género • Masculino. • Femenino. • Otro. • Prefiero no responder.
[02] ¿Cuál es el tipo de gestión de tu colegio? • Público. • Privado. • Privado con ayuda estatal.
[03] ¿En qué zona se encuentra tu colegio? • Norte. • Centro. • Sur. • Este. • Oeste.
[04] El próximo año, ¿te vas a dedicar a… • Estudiar? [Continuar en 05] • Trabajar? [Fin de encuesta] • Tomar año sábatico? [Fin de encuesta]
[05] ¿Qué carrera piensas elegir?
[06] ¿Sabes de qué trata la carrera de Soporte y Mantenimiento Informático?
[07] Si te gustaría recibir más información indica un canal de comunicación.

Piensen en 2 o 3 preguntas que estarían dispuestos a responder. Las preguntas formarán parte de una encuesta que se aplicará a todos los estudiantes del curso para buscar "regularidades" entre todos los cursantes. Pueden ser de todo tipo de identificación, de información, de opinión, abiertas o cerradas, etc. (ver la lectura n. º 4). La base de datos que se conforme la usaremos para hacer análisis de datos en el TP n.º 4.