Probabilidad y Estadística. Unidad 1: La Estadística.

Año de cursado: 1°

Clase N.º 2: La Estadística, definiciones y conceptos básicos.

Contenido: Tablas de frecuencias simples y por intervalos. Representación gráfica. Clasificación de las distribuciones de frecuencias.

1. Presentación:

Bienvenidos y bienvenidas a la clase n°2 de Probabilidad y Estadística. En el encuentro anterior comenzamos a transitar el espacio precisando algunas definiciones acerca de la Estadística como rama de la matemática, algunos de sus alcances como herramienta de investigaciones y, también, especificaciones y funcionalidades para el manejo de datos de diferentes naturalezas.

En esta clase, continuaremos avanzando sobre aspectos vinculados con la presentación de datos cualitativos y cuantitativos. En este sentido, trabajemos en particular sobre la confección de tablas de frecuencias, las cuales nos permitirán organizar los datos obtenidos para, posteriormente, avanzar sobre la etapa de síntesis que habíamos descrito en las etapas del método estadístico.

Trabajaremos, además, sobre los gráficos asociados a las tablas de frecuencias, así como también comenzar a observar algunos comportamientos de la forma que adquieren o más precisamente se distribuyen esos datos que hemos obtenidos al observar el fenómeno de estudio. Comencemos entonces.

2. Desarrollo:

Retomemos en primer lugar, la clasificación de variables. Habíamos dicho en la primera clase que las variables podrían clasificarse en dos grupos: **cualitativas y cuantitativas**. En ese sentido, para poder identificar las variables nos puede ayudar preguntarnos si los datos son o no valores numéricos, si la respuesta es sí, entonces la variable es cuantitativa; en caso contrario, la variable es cualitativa. Ahora bien, avanzando un poco más, dentro del grupo de las variables cuantitativas encontramos dos subgrupos y aquí una pregunta que nos podría ayudar a identificar con qué tipo de variable estamos trabajando seria: los datos ¿provienen de un proceso de conteo o de un proceso de medición? En el caso de la respuesta sea que los datos provienen de un proceso de conteo, diremos que la variable es **cuantitativa discreta**, mientras que cuando la respuesta sea que provienen de un proceso de medición (porque se hace uso de un instrumento de medición para tomar el dato), diremos que la variable es **cuantitativa continua**.

Con respecto a las variables cualitativas, para poder identificar la subclasificación que corresponde a este grupo, la pregunta es ¿se pueden ordenar según la importancia de cada valor de la variable? Si la respuesta es no, la variable es **cualitativa no ordenable**, en caso contrario la variable será **cualitativa ordenable**. En este punto vale una pequeña aclaración, aunque las variables cualitativas no ordenables no se organizan siguiendo un orden jerárquico, muchas veces se recurre a un ordenamiento alfabético que no implica jerarquía, pero si organiza el conjunto de datos, y recuerden la ida en este punto es poder organizar los datos de alguna manera.

Hacemos esta distinción porque cuando construimos tablas distribución de frecuencias, debemos tener en cuenta la clasificación de la variable para organizar los datos. Antes de continuar, vamos a definir qué es una tabla distribución de frecuencia, entonces, *es una forma de organizar los datos obtenidos en el proceso de recolección asignándole un valor que refiere a la cantidad de veces que aparece este dato en el conjunto sobre el cual estamos trabajo*. Esa cantidad de veces (ocurrencia) es lo que denominamos **frecuencia**.

Ahora, trabajaremos sobre las variables cualitativas. Para poder organizarlas en tablas de distribución frecuencias[1] vamos a considerar dos tipos:

1. **Tablas de distribución frecuencias simples:** son aquellas tablas que utilizamos para variables cuantitativas discretas, y es muy similar a la que se usa para poder organizar los datos de una variable cualitativa o categórica.
2. **Tablas de distribución de frecuencias con intervalos de clase:** son aquellas tablas que contienen para su organización intervalos de clase[2]. Se suelen usar para organizar datos que correspondan a una variable cuantitativa continua o variable discreta de amplio recorrido. Precisaremos más al respecto cuando trabajemos sobre esta tabla en particular.

[1] Muchas veces para simplificar nos referimos a las tablas de distinción de frecuencias como tablas de frecuencias, aunque la primera expresión corresponde a una definición y lenguaje más preciso.
[2] Un intervalo en matemática es un conjunto de valores que se encuentra entre dos valores fijos llamados extremos por lo general, vamos a utilizar la notación [a;b) para referirnos al conjunto de valores que se encuentra desde "a" hasta antes de "b". En este caso se incluyen valores muy cercanos "b" pero no al mismo "b", no se asusten vamos a retomarlo más adelante.

Comencemos a ver entonces las características de estas tablas con las que vamos a trabajar.

Tablas de distribución de frecuencias simples

Esta tabla tiene varios componentes que definiremos a continuación, pero antes me gustaría presentarla por medio de un ejemplo.

- Se tomaron datos correspondientes a los casos notificados de enfermedades de adicciones en 28 centros de salud de la provincia de Tierra del Fuego AIAS.
  5 6 7 5 6 12 6 9 5 7 5 6 7 8
  9 7 7 6 7 7 11 10 10 9 8 8 5 7

Para este caso podemos decir que la población son todos los centros de salud de la provincia, mientras que la muestra seria esos 28 centros de salud seleccionados. La variable son la cantidad de casos de adicción notificados.

Ahora vamos a organizarlos en una tabla de distribución de frecuencias simples. Para ello, en primer lugar, vamos a identificar el menor valor de la variable que es 5 y el mayor de la variable que para nuestro caso es el 12. Posteriormente, listamos de menor a mayor escribiendo cada valor una sola vez en la primera columna de nuestra tabla.

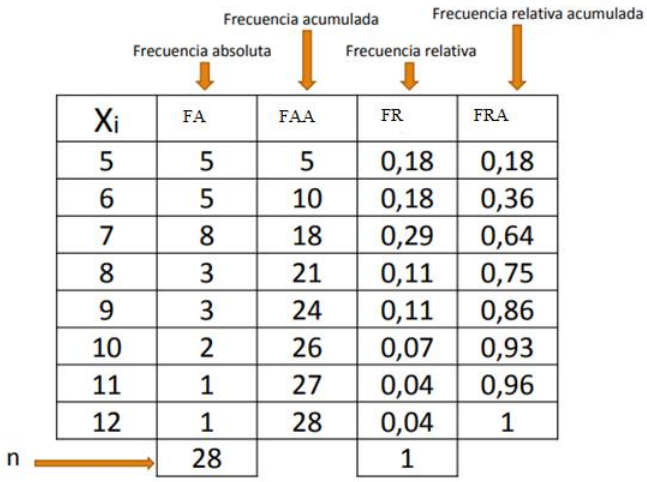
	X_i	FA	FAA	FR	FRA
Valores de la variable	5				
	6				
	7				
	8				
	9				
	10				
	11				
	12				

Seguidamente, vamos a precisar qué son cada uno de las notaciones que parecen como título de cada uno de las columnas.

1. FA: denominaremos de esta manera a las **frecuencias absolutas** y para completar esta columna debemos contar cuantas veces aparece cada uno de los valores de la variable y colocarlas en la fila correspondiente, por ejemplo, el valor 5 aparece 5 veces, el valor 6 aparece también 5 veces, y así con cada uno de los valores. La suma de todas las frecuencias acumuladas es igual al total de observaciones que en este caso es 28.
2. FAA: con esta notación nos referimos a las **frecuencias absolutos acumuladas**, y para completar esta columna, se comienza poniendo en la primera fila de esta columna el mismo valor que corresponde a FA, y para los siguientes debemos sumar las frecuencias absolutas anteriores hasta la fila correspondiente al valor de la variable que deseamos completar, por ejemplo, para completar la fila que corresponde al 6 en la columna de FAA, tenemos que sumar 5 más la fi que le corresponde al 6, es decir, 5 nuevamente, entonces 5+5=10 que sería el valor de la fa para el 6. La fa que corresponde al mayor valor de la variable debe coincidir con la suma de las FAA.
3. FR: utilizamos esta notación para referirnos a las **frecuencias relativas**. Para completar esta columna, debemos dividir la frecuencia absoluta correspondiente al valor de la variable sobre la cual vamos a completar la columna de FR por el total observaciones que para nuestro caso es 28. Por ejemplo, la fr que le corresponde al 5 es 5:28. La suma de todas las fr nos debe dar como resultado 1.

4. FRA: denominamos de esta manera a las frecuencias relativas acumuladas, para completar esta coluna se procede de manera análoga a la que utilizamos para completar las frecuencias acumuladas con la diferencia de que vamos a sumar FR.

Entonces, nuestra tabla quedaría de la siguiente manera:



X_i	FA	FAA	FR	FRA
5	5	5	0,18	0,18
6	5	10	0,18	0,36
7	8	18	0,29	0,64
8	3	21	0,11	0,75
9	3	24	0,11	0,86
10	2	26	0,07	0,93
11	1	27	0,04	0,96
12	1	28	0,04	1
	28		1	

Como podrán notar, aparece la letra n, habíamos dicho en la clase anterior que de esta manera denotábamos al tamaño de la muestra.

En algunos casos, se suelen incluir en las tablas las frecuencias relativas porcentuales que serían el resultado de multiplicar fr*100, entre otras columnas más.

Este desarrollo, corresponde a un trabajo que por lo general se hace de manera manual, pero para los fines de este curso vamos a utilizar el software que les había mencionado la clase pasada: InfoStat. La tabla construida en con el programa quedaría de la siguiente manera:

Tablas de frecuencias

Variable	Clase	FA	FR	FAA	FRA
Columnal	1	5	0,18	5	0,18
Columnal	2	5	0,18	10	0,36
Columnal	3	8	0,29	18	0,64
Columnal	4	3	0,11	21	0,75
Columnal	5	3	0,11	24	0,86
Columnal	6	2	0,07	26	0,93
Columnal	7	1	0,04	27	0,96
Columnal	8	1	0,04	28	1,00

En el siguiente link vamos a encontrar como realizar una tabla de distribución de frecuencias simples.

Ahora que ya hemos trabajado con las tablas de distribución de frecuencias simples avancemos un poco más, veamos entonces las tablas de distribución de frecuencias con intervalos de clase.

Tablas de distribución de frecuencias con intervalos de clase

Como podemos advertir, al trabajar con datos que corresponden a una variable cuantitativa continua (tomaremos ese caso primero y luego se precisará de las variables discretas de amplio recorrido) los valores que puede asumir la variable son números reales, es decir no solamente son números de contar (1;2; 3;) sino que también son valores numéricos decimales o irracionales (1,12; -0,54; …) resultaría impráctico confeccionar una tabla de similares características a la anterior, es por ello que para construir una tabla cuando la variable es continua utilizamos intervalos, es decir distribuimos los valores en grupos iguales que denominamos intervalos y se contabilizan los valores que se encuentran dentro de dicho intervalo. Veamos un ejemplo práctico para que se comprenda mejor.

Los siguientes datos corresponden a un conjunto de mediciones realizadas en una planta petrolera respecto del peso en kg de los barriles de petróleo producidos en un día.

134,574 134,504 133,479 135,994 133,093 135,459 133,846 138,241
137,587 134,649 134,542 135,507 134,323 136,810 134,966 135,215
134,957 131,829 136,662 135,086 135,647 134,324 137,089 134,857
136,007 135,018 135,323 134,558 136,021 134,869 138,113 132,973
137,445 135,116 134,667 136,179 134,731 133,140 134,94 136,522
134,340 135,409 134,323 136,280 134,333 134,138 132,934 136,421
134,767 133,701

Como podemos ver, los datos asumen diferentes valores que debemos organizar por intervalos de clase porque hay valores que se asemejan bastante pero no coinciden.

El primer paso para construir una tabla de frecuencia seria identificar el mínimo valor que toma la variable que para nuestro caso es 131,83 mientras que el máximo valores es 138,24. Vamos a calcular el **Rango**, para ellos hacemos la diferencia entre el máximo y el mínimo valor de la variable, es decir $138{,}241 - 131{,}829 = 6{,}412$

Seguidamente, vamos a utilizar el rango para determinar la **amplitud del intervalo**[3]. Para poder realizar el cálculo debemos definir la **cantidad de intervalos** (suele nombrarse a la cantidad de intervalos con la letra "k") que tendrá nuestra tabla, es recomendable que la cantidad de intervalos sea mayor a 5 para no perder precisión y menor a 15 para que la tabla no resulte

[3] La amplitud del intervalo se suele nombrar con la letra "h", e indica la diferencia entre los valores extremos que definen al intervalo, tranquilos ya cuando avancemos con el ejemplo será todo más claro.

impráctica. Entonces, para nuestro ejemplo vamos a determinar que la cantidad de intervalos sea 7. Ahora bien, tomamos el rango que habíamos calculado anteriormente y lo dividimos por 7, o sea $6{,}412 : 7 = 0{,}916$, lo cual quiere decir que la amplitud de cada intervalo será de 0,916. Armemos el primer intervalo, para ello, tomamos el menor valor que se denominará **límite inferior del intervalo de clase (LI)**, que en nuestro caso es 131,829 y sumémosle la amplitud (0,916), de esta manera obtenemos el **límite superior del intervalo de clase (LS)** que será 132,745, y de esta manera tenemos armado nuestro primer intervalo que los escribiremos: [131,241; 132,745).

Para armar el segundo intervalo, tomamos como límite inferior del intervalo el LI del intervalo anterior que es 132,745 y le sumamos nuevamente la amplitud 0,916 y obtenemos como LS del segundo 133,661, entonces nuestro segundo intervalo será [132,745; 133,661), y así proseguimos hasta que nuestro máximo valor quede contenido en el último intervalo.

En este punto valen varias aclaraciones, el último intervalo de nuestra tabla será [137,325;138,241], para este caso se coloca al final de intervalo un corchete que nos va a indicar que el mismo incluye al valor 138,241. En otros ámbitos se suelen hacer alguno tipo de salvedad para que esto no ocurra, pero para las finalidades de este curso no vamos a profundizar en ello puesto que trabajaremos con el software y el mismo incluye las consideraciones necesarias.

Para no extendernos demasiado, vamos a ver cómo nos queda la tabla en InfoStat:

Tablas de frecuencias

Variable	Clase	LI	LS	MC	FA	FR	FAA	FRA
Peso en kg de barriles de ..	1	[131,83	132,75)	132,29	1	0,02	1	0,02
Peso en kg de barriles de ..	2	[132,75	133,66)	133,20	5	0,10	6	0,12
Peso en kg de barriles de ..	3	[133,66	134,58)	134,12	12	0,24	18	0,36
Peso en kg de barriles de ..	4	[134,58	135,49)	135,04	16	0,32	34	0,68
Peso en kg de barriles de ..	5	[135,49	136,41)	135,95	7	0,14	41	0,82
Peso en kg de barriles de ..	6	[136,41	137,33)	136,87	5	0,10	46	0,92
Peso en kg de barriles de ..	7	[137,33	138,24]	137,78	4	0,08	50	1,00

Entonces, en esta tabla nos encontramos con una nueva columna desconocida **MC**, lo cual significa marca de clase, y no es más que la semisuma del LI y LS de cada intervalo, es decir límite inferior más límite superior y todo ello dividido por dos. Este valor será de mucha utilidad para el trabajo que desarrollaremos en la unidad 2.

Para comprar la columna que corresponde a FA, es decir lo que antes habíamos definido como frecuencia absoluta e indicaba cuantas veces se repetía cada valor de la variable, para este tipo de tablas nos indica la cantidad de números que están comprendidos en cada intervalo. Es decir, para el primer intervalo tenemos un solo valor comprendido desde 131,83 y el número inmediatamente anterior a 132,75, es decir, en el primer intervalo no se cuenta el LS, recién se lo cuenta en el segundo intervalo.

Dato curioso, como se habrán dado cuenta, el software solo trabajó con dos dígitos, no hay problema porque ya tiene incorporada la función y nos simplifica el trabajo.

Por último, cuando dijimos anteriormente se utilizaban tablas de distribución de frecuencias para organizar datos de variables discretas de amplio recorrido, nos referimos al caso en los que los valores parecen muy dispersos

y resultaría impráctico organizarlos en una tabla de distribución simple, la comprobación de ello lo realizáramos como actividad de la clase.

Para cerrar les dejo como material complementario, los tipos de gráficos asociados a cada una de las tablas y link de cómo construirlos en InfoStat porque nos hemos extendido bastante.

P/D: Para las tablas que corresponden a las variables cualitativas van a poder encontrarlo en el primer link, sigue principios análogos a los que hemos trabajado con estas dos tablas.

Les dejo el siguiente link donde se especifica como armar una tabla de distribución de frecuencias en InfoStat: https://youtu.be/LAbORsRruHY

Gráficos en InfoStat: https://youtu.be/NTZ6C4078FE

Actividad:

Construir en InfoStat una tabla de distribución de frecuencias simples y otra con 7 intervalos. Para cada caso, realizar los gráficos correspondientes.

Los siguientes valores corresponden a las edades de personas que sufrieron lesiones en accidentes de tránsito en el último año. El estudio fue realizado en el marco de un programa provincial de prevención vial, y las personas encuestadas son personas encuestadas habitantes de la provincia.

38 15 10 12 62 46 25 56 27 24 23 21 20 25 38 27 48 35 50 65
59 58 47 42 37 35 32 40 28 14 12 24 66 73 72 70 68 65 54 48
34 33 21 19 61 59 47 46 30 30 2 6 10 16 20 26 27 29 24 24 20
20 21 17 16 16 19 18 3 2 12 12 10 10 10 14 14 13 12 5 9 514 10 14 13 11
10 9 8 5 6 7 29 28 31 30

- Indicá la población, muestra y variable (clasifica)

Para la entrega de las actividades dejo habilitado un espacio de tareas donde podrán realizar la entrega correspondiente. Para ello, les sugiero que armen un archivo en Word y peguen allí las imágenes de las producciones que realicen en InfoStat y también puedan registrar la población, muestra y variable con su correspondiente clasificación.

3. Actividad integradora de cierre:

Les proponemos analizar la siguiente situación:

El Ministerio de Salud de la Nación, en colaboración los Ministerios de Educación Nacional y Provinciales, la OMS Argentina, OMS Washington y los centros para el control y la Prevención de Enfermedades (CDC), llevó a cabo la tercera Encuesta Mundial de Saludos Escolar (EMSE 2018) en Argentina. Este estudio tuvo como objetivo recopilar datos precisos sobre comportamientos relacionados con la saludo y los factores de riesgos y protección entre estudiantes de 13 a 17 años en el país.

En esta ocasión, encuentras a cargo del tratamiento de los datos extraídos de la EMSE 2018, a los cuales podés acceder en el siguiente enlace: https://acortar.link/16Gytq[4]. Tu tarea consiste en identificar la población, muestra, variable con su correspondiente clasificación, elaborar tablas de frecuencias y seleccionar datos relevantes para su interpretación. Se espera que realices la presentación de un informe considerando todos los puntos

[4] Elaboramos una base de datos, a partir de la fuente EMSE disponible en: http://datos.salud.gob.ar/dataset/base-de-datos-de-la-3-encuesta-mundial-de-salud-escolar-emse-con-resultados-nacionales-argentina

mencionados anteriormente y aquellos no mencionados pero que consideres relevante.

Se habilitará un espacio el campus para la entrega de modo que pueda quedar registro de esta actividad, aunque se prevé realizar una socialización y corrección de la actividad en el encuentro sincrónico correspondiente a la clase nro. 2.

4. Cierre:

Para resumir, en esta clase hemos avanzado sobre la clasificación de las variables y profundizamos sobre la construcción de tablas para diferentes tipos de variable, así como también los gráficos asociados a cada una de las tablas.

Dejo abierto el foro para consultas.

Saludos, los y las espero en la próxima clase.

5. Bibliografía:

- García, J; López, N; Calvo, J. (2011). Estadísticas Básicas para Estudiantes de Ciencias. Facultad de Ciencias Físicas Universidad Complutense de Madrid. España.
- Wackerly, D; Mendenhall, W; Schaeffer, R. (2010). Estadística Matemática con Aplicaciones. 7ma Ed. Cengage Learning. Santa Fe, México.