



Profesora Teresita Fuster Bardier
Licenciada en Estadística

Documento realizado para docentes de CETP
Montevideo, agosto de 2008

ÍNDICE GENERAL	1
1 CONCEPTOS GENERALES	4
1.1 ESTADÍSTICA	4
1.2 ESTADÍSTICA DESCRIPTIVA	4
1.3 ALGUNAS DEFINICIONES BÁSICAS	4
1.4 - VARIABLES ALEATORIAS	4
1.4.1 Concepto y ejemplos.....	4
1.4.2 Clasificación.....	4
2 ESTADÍSTICA DESCRIPTIVA	6
2.1 ANÁLISIS UNIVARIADO	6
2.1.1 Distribución de frecuencia y frecuencia acumulada	6
2.1.2 Representaciones Gráficas.....	6
2.1.2.1 Gráfica circular	7
2.1.2.2 Histograma	9
2.1.2.3 Gráfica de líneas	9
2.1.3 Medidas de Tendencia Central	10
2.1.3.1 Modo	10
2.1.3.2 Mediana	10
2.1.3.3 Media (o Promedio).....	11
2.1.4 Medidas de Dispersión.....	11
2.1.4.1 Rango.....	12
2.1.4.2 Varianza (o variancia).....	12
2.1.4.3 Desvío estándar	13
2.2 ANÁLISIS CONJUNTO DE VARIABLES	14
2.2.1 Tablas de contingencia	14
2.2.2 Gráficos	16
3. PROBABILIDAD	18
3.1 CONCEPTO Y AXIOMAS	18
2.1.1 Definiciones	18
2.1.3 Definición axiomática.....	19
2.2 PROPIEDADES	19
2.2.1 Propiedades generales.....	19
2.2.2 Probabilidad clásica	20
2.2.3 Probabilidad condicional	20
2.2.4 Sucesos independientes.....	21
2.2.5 Teorema de Bayes	21
3.3 DIFERENTES ENFOQUES	23
3.4 DISTRIBUCIONES DE VARIABLES ALEATORIAS	24
3.4.1 Variables discretas	24
3.4.1.1 Distribución de Bernoulli	24
3.4.1.2 Distribución binomial	24
3.4.1.3 Distribución de Poisson.....	25
3.4.2 Variables continuas.....	25
3.4.2.1 Distribución Normal.....	25
3.4.2.2 Distribución Exponencial.....	26
3.4.2.3 Otras distribuciones	26

4 INFERENCIA ESTADÍSTICA	27
4.1 INTRODUCCIÓN	27
4.2 ESTIMACIÓN PUNTUAL	27
4.3 ESTIMACION POR INTERVALO	27
4.4 PRUEBAS DE HIPÓTESIS	28
4.5 TEST DE INDEPENDENCIA	28
4.5.1 Variables cualitativas	28
3.5.2 Variables cuantitativas	29
4.1 CONCEPTOS BÁSICOS	30
4.1.1 Población.....	30
4.1.2 – Encuesta por muestreo	30
4.1.3 – Marco muestral	30
4.1.4 – Muestreo probabilístico	30
4.2 DIFERENTES DISEÑOS DE MUESTREO	31
4.2.1 Muestreo Aleatorio Simple.....	31
4.2.2 Muestreo sistemático	31
4.2.3 Muestreo estratificado.....	32
4.2.4 Muestreo por conglomerados y en varias etapas	32
5 TÉCNICAS MULTIVARIADAS	33
5.1 MODELOS DE REGRESIÓN LINEAL	33
5.2 MODELOS DE REGRESIÓN LOGÍSTICA	33
5.3 ANÁLISIS FACTORIAL	34
5.3.1 Análisis de Componentes Principales (ACP).....	34
5.3.2 Análisis de Correspondencia Simple (ACS).....	34
5.3.3 Análisis de Correspondencia Múltiple (ACM)	35
5.4 ANÁLISIS DE CLUSTERS	35
5.5 ANÁLISIS DISCRIMINANTE	35
BIBLIOGRAFÍA	36

1 CONCEPTOS GENERALES

1.1 ESTADISTICA

Es la ciencia que tiene por objeto la recolección, la organización, el análisis y la presentación de datos, con el fin de brindar información que facilite la toma de decisiones.

La Estadística es una rama del conocimiento particularmente nueva y de gran aplicabilidad. Aunque sus orígenes se remontan al siglo XIX, con los estudios sobre antropometría del belga Adolphe Quetelet y sobre herencia del inglés Francis Galton, muchos de sus avances se han dado a partir de mitad de siglo XX, en parte gracias al progreso de la informática.

La utilización de los métodos estadísticos va desde la Medicina a la Economía, pasando por la Agronomía, las Ciencias Sociales, las Ciencias Políticas y el Marketing.

1.2 ESTADÍSTICA DESCRIPTIVA

Es el primer paso en cualquier análisis estadístico. Resume los datos en porcentajes o números que son fácilmente interpretables y comparables con otros datos similares. Así mismo pueden proporcionarse gráficas que resuman estos datos.

1.3 ALGUNAS DEFINICIONES BÁSICAS

a) Población: Una *población finita* (o simplemente población) es un conjunto finito de elementos. Esto implica que puede determinarse sin ambigüedad si un elemento pertenece o no al conjunto. Denominaremos **U** a nuestra *población objetivo*, o sea la población finita sobre la cual deseamos obtener alguna información estadística.

b) Muestra: Es un subconjunto de la población. En algunas ocasiones no tenemos los datos totales de la población, sino solamente de una parte de ella. En estos casos, las conclusiones extraídas de estos datos pueden extenderse o no a la población dependiendo de cómo se haya extraído la muestra.¹

1.4 - VARIABLES ALEATORIAS

1.4.1 Concepto y ejemplos

Una variable es cualquier dato sujeto a medida o cuenta. Es aleatoria si no se puede predecir su valor.

Ejemplos:

- ✓ La hora de salida del sol cada día no es una variable aleatoria, ya que los astrónomos saben de antemano la hora exacta de la misma para cada día del año.
- ✓ La cantidad de lluvia caída durante un período específico sí es una variable aleatoria, ya que no puede predecirse.
- ✓ La cantidad de alumnos inscriptos en determinado curso en el año próximo.
- ✓ El número de alumnos que promueven el curso en determinada orientación.

1.4.2 Clasificación

Las variables aleatorias suelen clasificarse según la naturaleza de los datos a los que se refieran:

¹ Se ampliará este concepto en la unidad correspondiente a Muestreo Aleatorio.

Introducción a la Probabilidad y Estadística

- i- **Cualitativas:** También se llaman categóricas. Los datos están divididos en clases o categorías. Por ejemplo: el tipo de curso; sexo del alumno, orientación, etc.
- ii- **Cuantitativas:** Son datos numéricos. Se diferencian en
 - a) Discretas: son aquellos datos que se pueden contar o numerar. Ejemplos: edad del alumno, cantidad de alumnos matriculados en cada escuela; cantidad de profesores de Taller por departamento; cantidad de horas libres que tiene el Centro por día
 - b) Continuas: datos numéricos que pueden tomar cualquier valor (en general, entre ciertos límites). Por ejemplo: salario por hora docente en el tiempo o en diferentes grados.

2 ESTADÍSTICA DESCRIPTIVA

2.1 ANÁLISIS UNIVARIADO

En el siguiente apartado se trabajará con las diferentes formas de presentar los datos referidos a una única variable.

2.1.1 Distribución de frecuencia y frecuencia acumulada

La frecuencia de una variable (tanto discreta como cualitativa) es la cantidad de veces que el dato se repite. Normalmente los datos se presentan agrupados según una tabla de frecuencias, que puede contener frecuencias absolutas (número de casos) o frecuencias relativas (porcentajes)

En el ejemplo siguiente se muestra la cantidad de alumnos de Enseñanza Media en Uruguay, dependiendo del tipo de curso en el que estaban inscriptos en el año 2006.

Tipo de Curso	Frecuencia	Porcentaje
Ciclo Básico CES	115.756	39,52
Ciclo Básico Privado	20.991	7,17
Ciclo Básico Tecnológico UTU	16.822	5,74
Formación Profesional UTU	13.645	4,66
Formación Profesional Privado	1.257	0,43
Bachillerato Diversificado CES	89.464	30,54
Bachillerato Diversificado Privado	17.351	5,92
Bachillerato Tecnológico UTU	17.651	6,03
Total	292.937	100,00

Cuadro N° 1 – Alumnos de Enseñanza Media según tipo de curso
Fuente: Encuesta Nacional de Hogares Ampliada – INE 2006

2.1.2 Representaciones Gráficas

Los datos pueden ser también presentados mediante gráficas, cuya confección depende del tipo de variable con la que se esté trabajando.

2.1.2.1 Gráfica de barras

Se utilizan para variables cualitativas.

Consisten en tantos rectángulos como categorías tiene la variable en cuestión. Las bases de estos rectángulos deben ser iguales. La altura es proporcional a la frecuencia de la categoría (o sea, a la cantidad de veces que se encuentra el dato)

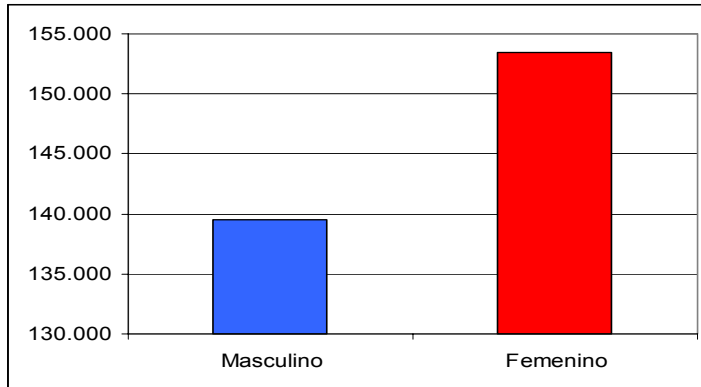
Ejemplos:

- a) Usando los datos vistos en el caso anterior, la distribución por género de estos alumnos es la siguiente:

Género	N° alumnos
Masculino	139.521
Femenino	153.416
Total	292.937

Cuadro N° 2: Número de alumnos matriculados en Enseñanza Media según género
Fuente: Encuesta Nacional de Hogares Ampliada – INE 2006

La gráfica de barras correspondiente a estos datos se muestra a continuación

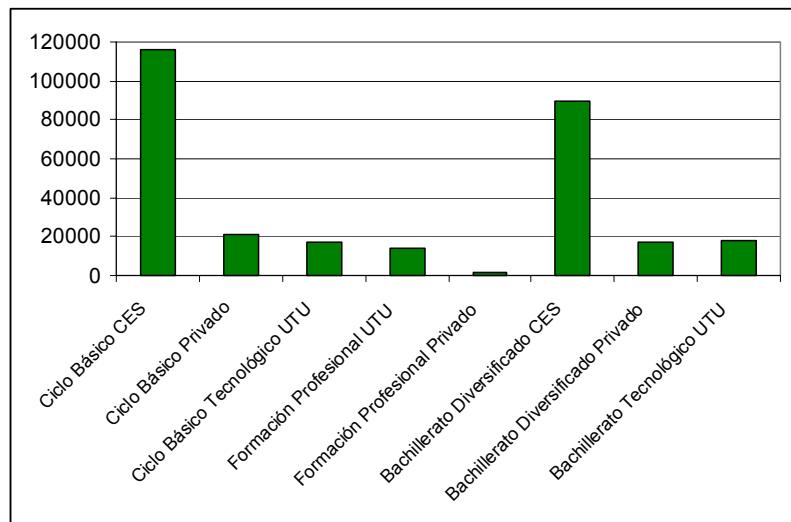


Gráfica N° 1: Número de alumnos según género

Fuente: Encuesta Nacional de Hogares Ampliada – INE 2006

b) Los datos vistos en el apartado anterior, pueden representarse también mediante una gráfica de barras:

Gráfica N° 2: Número de alumnos según tipo de curso
Fuente: Encuesta Nacional de Hogares Ampliada – INE 2006



2.1.2.2 Gráfica circular

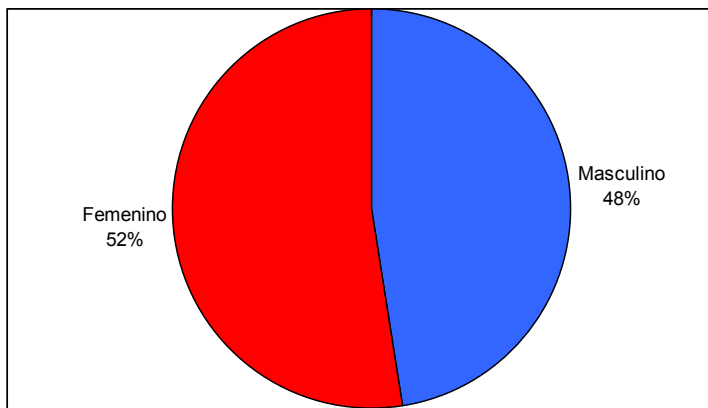
También se utilizan para variables cualitativas. En este caso, el total de datos se representan en un círculo. Éste se divide en tantos sectores como categorías tiene la variable. La amplitud (medida en grados) de cada uno de estos sectores es proporcional a la frecuencia de la categoría que representa.

a) En el ejemplo de la distribución de alumnos según género:

	Frecuencia	Ángulo
Masculino	139.521	171
Femenino	153.416	189
Total	292.937	360

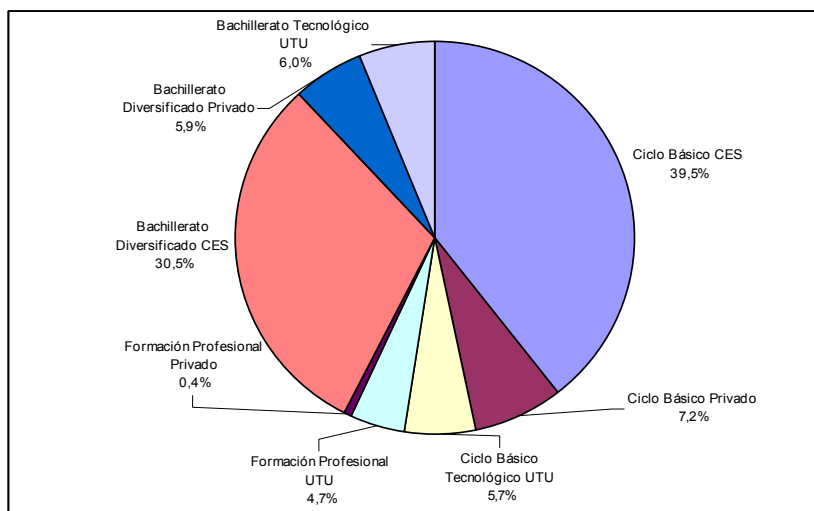
Cuadro N° 3: Número de alumnos matriculados en Enseñanza Media según género
Fuente: Encuesta Nacional de Hogares Ampliada – INE 2006

La gráfica correspondiente es:



Gráfica N° 3: Número de alumnos según género
Fuente: Encuesta Nacional de Hogares Ampliada – INE 2006

b) En el caso de la distribución de los alumnos por tipo de curso, la gráfica que se obtiene es la siguiente:



Gráfica N° 4: Número de alumnos según tipo de curso
Fuente: Encuesta Nacional de Hogares Ampliada – INE 2006

Como puede apreciarse en los ejemplos, la utilización de la gráfica circular es preferible cuando el número de categorías es relativamente pequeño, ya que de lo contrario puede perderse el efecto visual de la gráfica.

2.1.2.3 Histograma

El histograma es una gráfica de barras usada para variables cuantitativas discretas. La diferencia es que las barras son consecutivas (o sea, no hay espacios entre ellas).

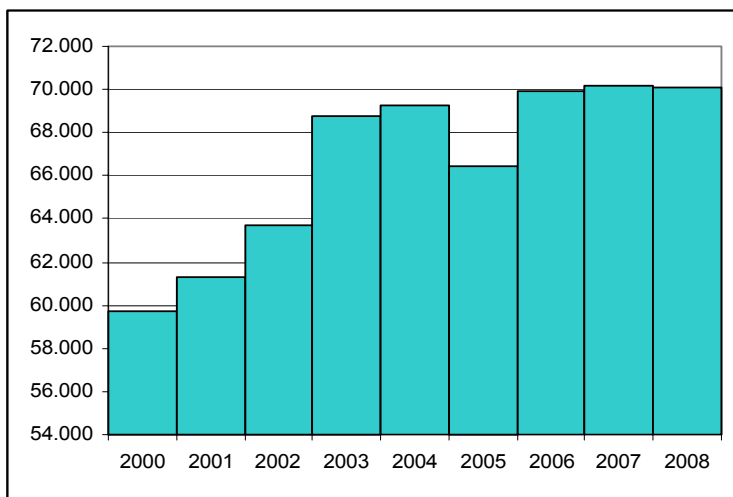
Ejemplo:

El número de alumnos matriculados en UTU desde el año 2000 al año 2008 ha sido el siguiente:

Año	Nº alumnos
2000	59.716
2001	61.327
2002	63.676
2003	68.779
2004	69.222
2005	66.429
2006	69.896
2007	70.184
2008	70.110

Cuadro N° 4: Número de alumnos matriculados en UTU según año

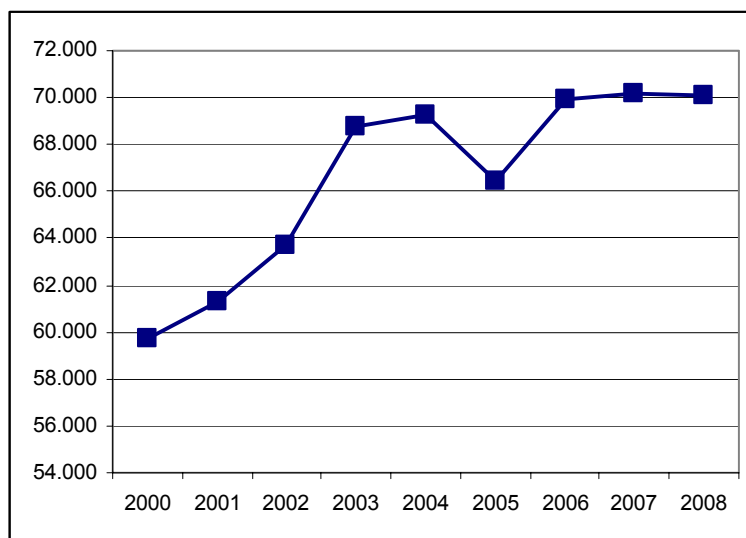
Fuente: Departamento de Estadística – Programa Planeamiento Educativo - CETP



Gráfica N° 5: Número de alumnos matriculados en UTU según año

2.1.2.4 Gráfica de líneas

Este tipo de gráficas es utilizado para variables cuantitativas, tanto discretas como continuas. En el caso de las variables discretas, se toma como valor en el eje horizontal el punto medio entre dos valores consecutivos. También se les conoce como polígonos de frecuencias. En el ejemplo anterior:



Gráfica N° 6: Número de alumnos matriculados en UTU según año

2.1.3 Medidas de Tendencia Central

En la mayoría de los casos de la vida real, la cantidad de datos es abrumadora. Es imposible poder manejarlos en su totalidad y tampoco se pueden efectuar comparaciones entre series diferentes de la misma variable. Para evitar estos problemas, se suelen utilizar ciertos valores que resumen esos datos. En general se denominan Medidas de Tendencia Central. Se estudiarán los tres más usados: modo, mediana y media.

2.1.3.1 Modo

Se puede utilizar en variables *cuantitativas o cualitativas*. También se le conoce como “Moda”.

Es el valor de la variable que presenta la mayor frecuencia. Pueden presentarse ejemplos en los cuales dos valores de la variable presentan igual frecuencia (o muy semejantes, si son números grandes). En estos casos se dice que la variable tiene una *distribución bimodal*.

Si se observa el ejemplo de la distribución de alumnos matriculados en Educación Media según el tipo de curso, la modalidad que tiene mayor frecuencia es “Ciclo Básico CES”, por lo cual es el modo de la distribución.

En el ejemplo de la distribución de estos alumnos por género, el modo es la categoría “Femenino”.

En el caso de la cantidad de alumnos matriculados en el CETP durante el período 2000-2008, el modo, estrictamente hablando es “2007”, aunque también puede considerarse que la matrícula es prácticamente constante en los últimos tres años.

2.1.3.2 Mediana

Se utiliza para variables *cuantitativas*.

Es el valor de la variable que se encuentra en el lugar central una vez ordenados los datos en forma creciente. O sea, la mitad de los datos son menores que la mediana y la otra mitad, mayores.

Ejemplo:

Los alumnos de determinado grupo tiene las siguientes edades: 15, 15, 18, 17, 19, 16, 18, 16, 20, 18, 15, 16, 18, 20, 17

Al ordenarlos en forma creciente se obtiene:

15, 15, 15, 16, 16, 16, 17, 17, 18, 18, 18, 18, 19, 20, 20

El total de datos es 15, por lo tanto el centro de la distribución ordenada lo ocupa el dato que se encuentra en el 8° lugar, que en este caso es **17**. O sea, la mediana de esta distribución es 17.

Si el número de datos es impar, no hay dudas en cual es el lugar central de la distribución, ya que queda igual cantidad de datos a ambos lados de la mediana.

Si el número de datos es par, hay dos valores centrales. Para hallar la mediana se suman estos datos y el resultado se divide entre dos.

Cuando los datos son numerosos, es prácticamente imposible trabajar de esta manera. Si la variación de los datos no es muy grande, habrá valores repetidos. Estos datos se los agrupa en una tabla de frecuencias y luego se encuentran las frecuencias acumuladas (o sea, cuántos datos son menores o iguales a un determinado valor de la variable). Por último se busca que valor de la variable ocupa el lugar central. Si la variación es importante, se utilizan tablas de datos agrupados. En este caso, en lugar de trabajar con valores puntuales se trabaja con intervalos.

Existe también la posibilidad de realizar estos cálculos mediante software (por ejemplo, Planillas de Cálculo de Microsoft Office o de Open Office, así como los programas específicos para estadística).

2.1.3.3 Media (o Promedio)

Se utiliza para variables *cuantitativas*.

Es el valor que tomaría la variable si los datos fueran todos iguales, manteniendo el total. Se calcula sumando todos los datos y dividiendo entre el número de datos.

En el ejemplo visto anteriormente:

La suma de las edades de los alumnos es 242.

El total de alumnos considerados es 15.

Por tanto, la media es $242/15 = 16.13$ Media = 16.13

En este caso (al igual que en la mayor parte de los ejemplos reales) el valor de la media no es un valor posible de la variable, ya que se está trabajando con una variable discreta y el resultado es un número decimal.

Estos valores también pueden calcularse mediante programas adecuados (incluso muchas calculadoras científicas permiten ingresar datos y calcular la media)

Notas:

Cuando los datos se presentan por medio de alguna de las medidas de tendencia central, se pierden los valores particulares de los mismos, pero se gana en sencillez y practicidad.

La medida de tendencia central más utilizada es la media. Pero debe tenerse especial cuidado cuando en la variable existen valores muy alejados, por ejemplo, valores muy bajos y muy altos de la variable. Unos pocos datos (inclusive un único dato) con valores extremos pueden hacer aumentar considerablemente el valor de la media. Si se usa la mediana los valores extremos de la distribución no pesan tanto. Por eso es que pueden existir variaciones importantes entre las dos medidas.

2.1.4 Medidas de Dispersión

Cuando se comparan dos distribuciones de la misma variable, se utilizan en general las medidas de resumen, siendo la más frecuente, la media. Pero, en ciertas ocasiones, no alcanza con dar un solo valor.

Por ejemplo:

El número de alumnos matriculados en tres Institutos de enseñanza durante 5 años consecutivos, fueron los siguientes:

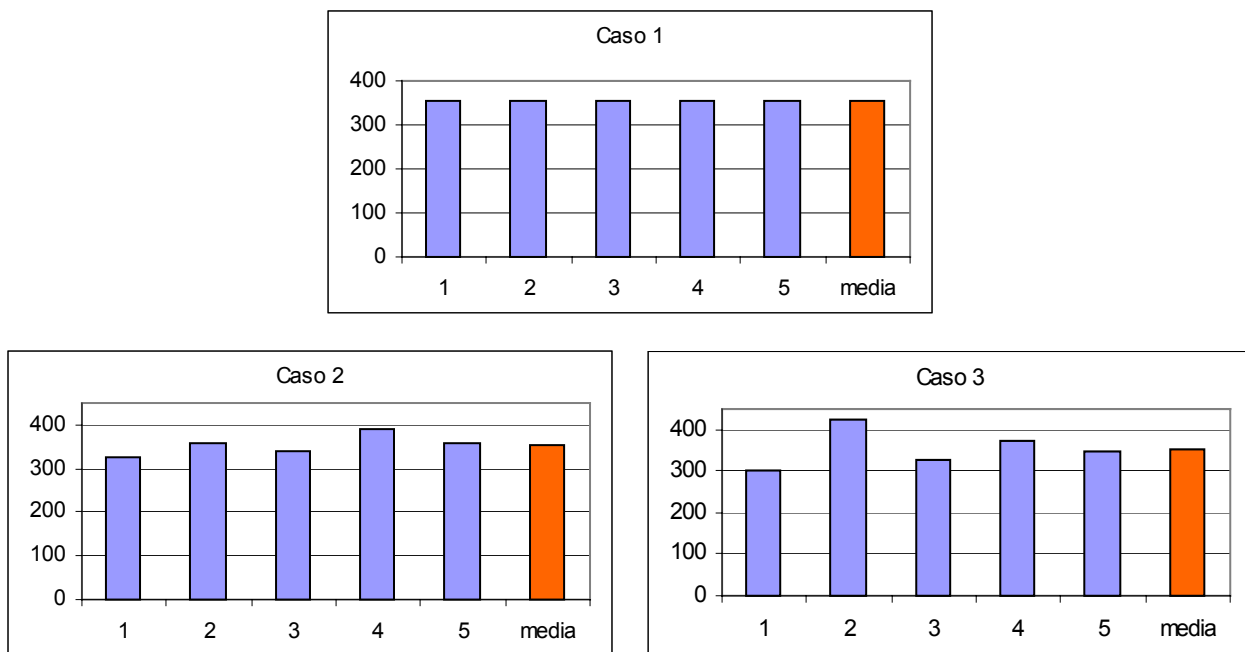
1^{er} caso: 355; 355; 355; 355; 355

2^o caso: 327; 360; 340; 390; 358

3^{er} caso: 300; 425; 328; 375; 347

Al efectuar los cálculos se observa que la media es 355 en los tres casos. Pero las distribuciones presentan diferencias marcadas: en la primera, todos los valores son iguales, lo cual no sucede en las otras dos. Si se observa la mediana, en el primer caso es 355, en el segundo es 358 y en el tercero es 347.

Si se observan las gráficas respectivas:



Gráfica N° 7: Diferentes distribuciones del número de alumnos matriculados
Fuente: creación propia

Una forma de comparar estas distribuciones es mediante las medidas de dispersión, entre las que se verán el rango, la varianza (o variancia) y el desvío estándar. En todos los casos se trabaja con variables numéricas, aunque los ejemplos se verán únicamente con variables discretas.

2.1.4.1 Rango

Es la amplitud de la distribución. Se calcula como la diferencia entre los valores extremos (o sea, el mayor y el menor)

En los ejemplos vistos:

1^{er} caso: el rango es 0 (todos los valores son iguales)

2^o caso: el rango es 63 (390 – 327)

3^{er} caso: el rango es 125 (425 – 300)

2.1.4.2 Varianza (o variancia)

Es una especie de promedio de las diferencias entre los valores observados y la media, elevados al cuadrado (porque de lo contrario, estas diferencias se compensan y el total da cero). La fórmula de cálculo es la siguiente: se hallan la diferencia de cada dato y la media y se eleva al cuadrado. Posteriormente se suman los valores obtenidos y este resultado se divide entre el número de datos menos 1.

En los ejemplos vistos

Caso 2

Dato	Media	Diferencia	Dif. al cuadrado
327	355	- 28	784
360	355	5	25
340	355	-15	225
390	355	35	1225
358	355	3	9
Suma		0	2268

Cuadro N° 5: Ejemplo de cálculo de la Variancia

La variancia es, entonces, 567 (obtenida al dividir 2268 entre 4)

Caso 3

Dato	Media	Diferencia	Dif. al cuadrado
300	355	- 55	3025
425	355	70	4900
328	355	- 27	729
375	355	20	400
347	355	-8	64
Sumas		0	9118

Cuadro N° 6: Ejemplo de cálculo de la Variancia

La variancia en este caso es 2279,5 (obtenida al dividir 9118 entre 4)

2.1.4.3 Desvío estándar

Debido a que para obtener la variancia es necesario elevar al cuadrado, la unidad de medida de la misma es el cuadrado de la original. Por este motivo, suele usarse la raíz cuadrada de la variancia, que se llama desvío o desviación estándar. En este caso, la unidad de medida de los datos se mantiene. Por ejemplo, si la variable es la altura de un grupo de estudiantes, la unidad de medida es el metro. Por los cálculos realizados, la unidad de medida de la variancia es el metro cuadrado, por lo que es más fácil de interpretar el número que determina el desvío estándar, ya que nuevamente la unidad de medida es el metro.

En los ejemplos vistos, el desvío estándar es 23,81 en el Caso 2 y 47,74 en el caso 3.

Es de destacar que en el primer caso, tanto la variancia como el desvío son ceros, ya que no hay variación entre los datos.

Cualquiera que sea la medida de dispersión usada, se observa que en el tercer caso, los datos están más dispersos que en el segundo caso y que en el primero. Cuanto más agrupados en torno a la media se encuentren los datos, menor será la medida de dispersión, cualquiera sea la que se usa.

2.2 ANÁLISIS CONJUNTO DE VARIABLES

En ocasiones, es necesario realizar el análisis de los datos vinculando dos o más variables. Si estas variables son cualitativas o, siendo numéricas se pueden modificar creando modalidades, la forma tradicional de tratarlas es mediante tabla de datos cruzados o Tablas de Contingencia. También pueden usarse gráficos que vinculen dos variables.

2.2.1 Tablas de contingencia

Como regla general, es conveniente que la tabla sea “más larga que ancha”. O sea, la variable que tiene más categorías debe presentarse en las filas y la de menos categorías en las columnas. Es simplemente un aspecto visual: ayuda en la interpretación de los datos. No es aconsejable que las tablas sean muy extensas ni incluir más de tres variables (siempre que dos de ellas tengan muy pocas categorías). Con estas tablas no se pretende, en general, que una variable explique a la otra, sino que la intención es ver como se corresponden las distintas modalidades de cada una de las variables (por ejemplo, si los datos que corresponden a una de las modalidades de una de las variables están concentrados o distribuidos en las modalidades de la otra variable).

Un ejemplo es el siguiente (también extraído de los datos de la Evaluación Diagnóstica)

Tramos de edad	Género		
	Total	Hombre	Mujer
Total	11327	6176	5151
Hasta 15 años	2402	1504	898
16 a 20 años	6433	3527	2906
21 a 25 años	1442	655	787
26 a 30 años	502	241	261
31 a 35 años	243	121	122
36 y más años	305	128	177

Cuadro N° 7: Número de alumnos por Género según Grupo de Edad
Fuente: Evaluación diagnóstica Nivel II – CETP - 2008

En estas tablas hay varias opciones: los totales pueden aparecer tanto al inicio de la tabla (más usados en las presentaciones internacionales) como al final. Se acostumbra a usar totales de filas y de columnas.

En este caso, la tabla se debe titular “Número de alumnos por sexo, según edad”.

En cuanto a la presentación de los datos de la tabla mediante porcentajes, depende del objetivo de la misma:

- porcentajes correspondientes a cada fila (o sea, el 100% es el número de observaciones de cada categoría de la variable que ocupa las filas. En el ejemplo: porcentaje de hombres hasta 15 años y porcentaje de mujeres hasta 15 años, etc.)
- porcentajes correspondientes a columnas (o sea, el 100% es el total de observaciones de cada modalidad de la variable que ocupa las columnas. En el ejemplo se tendría: del total de hombres, que porcentaje tiene cada grupo de edades)
- porcentaje de tabla (el 100% es el total de observaciones. En el ejemplo: porcentaje de hombres menores de 15 años, etc.)

Cuando la tabla se presenta con los totales de observaciones, la persona que esté leyendo el informe puede calcular las otras tablas según sus necesidades. Si solo se presentan los porcentajes, esto no es posible, salvo que se brinden los valores totales de las filas, las columnas o de tabla según corresponda.

Introducción a la Probabilidad y Estadística

Estas posibilidades se muestran en el siguiente ejemplo extraído también de la Evaluación Diagnóstica 2008. En este caso se vinculan las variables “Nivel socioeconómico”, como Estratos del Índice Socio Económico (INSE) y la variable resumen del resultado global de la prueba (con modalidades Aceptable y No aceptable)

- Número de alumnos por Promedio Global según Estrato del INSE

Estratos INSE	Calificación global		Total
	Insuficientes	Suficientes	
ALTO - ALTO	148	21	169
ALTO - MEDIO	760	99	859
MEDIO - ALTO	1962	183	2145
MEDIO - MEDIO	3364	221	3585
MEDIO - BAJO	2583	173	2756
BAJO - MEDIO	1625	115	1740
BAJO - BAJO	275	25	300
Total	10717	837	11554

Cuadro Nº 8: Número de alumnos por Calificación según Nivel Socio Económico
Fuente: Evaluación diagnóstica Nivel II – CETP - 2008

Lectura de la tabla: la casilla señalada indica que hay 1962 alumnos que se clasificaron como de nivel Socio – económico medio-alto y que obtuvieron promedio No aceptable en la Prueba aplicada.

- Porcentajes de filas

Estratos INSE	Promedio global		Total
	Insuficientes	Suficientes	
ALTO - ALTO	87,57	12,43	100,00
ALTO - MEDIO	88,47	11,53	100,00
MEDIO - ALTO	91,47	8,53	100,00
MEDIO - MEDIO	93,84	6,16	100,00
MEDIO - BAJO	93,72	6,28	100,00
BAJO - MEDIO	93,39	6,61	100,00
BAJO - BAJO	91,67	8,33	100,00
Total de tabla	92,76	7,24	100,00

Cuadro Nº 9: Porcentaje de alumnos por Calificación según Nivel Socio Económico
Fuente: Evaluación diagnóstica Nivel II – CETP - 2008

Lectura de la tabla: la casilla indica que del total de alumnos clasificados como de nivel socio-económico medio-alto, el 91.47% obtuvo promedio No aceptable en la Prueba aplicada.

- Porcentajes de columnas

Estratos INSE	Promedio global		Total
	Insuficientes	Suficientes	
ALTO - ALTO	1,38	2,51	1,46
ALTO - MEDIO	7,09	11,83	7,43
MEDIO - ALTO	18,31	21,86	18,56
MEDIO - MEDIO	31,39	26,40	31,03
MEDIO - BAJO	24,10	20,67	23,85
BAJO - MEDIO	15,16	13,74	15,06
BAJO - BAJO	2,57	2,99	2,60
Total	100,00	100,00	100,00

Cuadro N° 10: Porcentaje de alumnos por Calificación según Nivel Socio Económico
Fuente: Evaluación diagnóstica Nivel II – CETP - 2008

Lectura de la tabla: la casilla seña que del total de alumnos que tuvieron promedio insuficiente en la Prueba, el 18.31% se clasifican en el nivel Socio-económico medio-alto.

- Porcentajes de tabla

Estratos INSE	Promedio global		Total
	Insuficientes	Suficientes	
ALTO - ALTO	1,28	0,18	1,46
ALTO - MEDIO	6,58	0,86	7,43
MEDIO - ALTO	16,98	1,58	18,56
MEDIO - MEDIO	29,12	1,91	31,03
MEDIO - BAJO	22,36	1,50	23,85
BAJO - MEDIO	14,06	1,00	15,06
BAJO - BAJO	2,38	0,22	2,60
Total	92,76	7,24	100,00

Cuadro N° 11: Porcentaje de alumnos por Calificación según Nivel Socio Económico
Fuente: Evaluación diagnóstica Nivel II – CETP - 2008

Lectura de la tabla: la casilla señala que del total de alumnos que realizaron la Prueba diagnóstica, el 16.98% fueron clasificados en el nivel Socio económico medio-alto y obtuvieron promedio No aceptable.

Como indicación general, si una misma variable aparece en varias tablas, es conveniente que siempre se encuentre siempre como fila o como columna, ya que facilita la lectura de las tablas y las conexiones o comparaciones entre las mismas.

2.2.2 Gráficos

Cuando se necesita mostrar la información correspondiente a dos variables, se pueden usar gráficos de barras acumuladas: o sea, la barra correspondiente a cada modalidad de una de las variables se divide (en diferentes colores generalmente) para indicar las proporciones correspondientes a cada una de las modalidades de la otra variable.

Estos gráficos pueden confeccionarse tanto con valores absolutos (número de observaciones, cada gráfica tiene la altura proporcional al número de observaciones de la modalidad que representa) como con valores relativos (cada barra representa el 100%, por lo que todas tienen la misma altura)

Ejemplo: Distribución de los alumnos por edad y sexo

- El siguiente gráfico fue confeccionado con los totales de observaciones, por lo que también se observa la cantidad de alumnos correspondientes a cada modalidad.

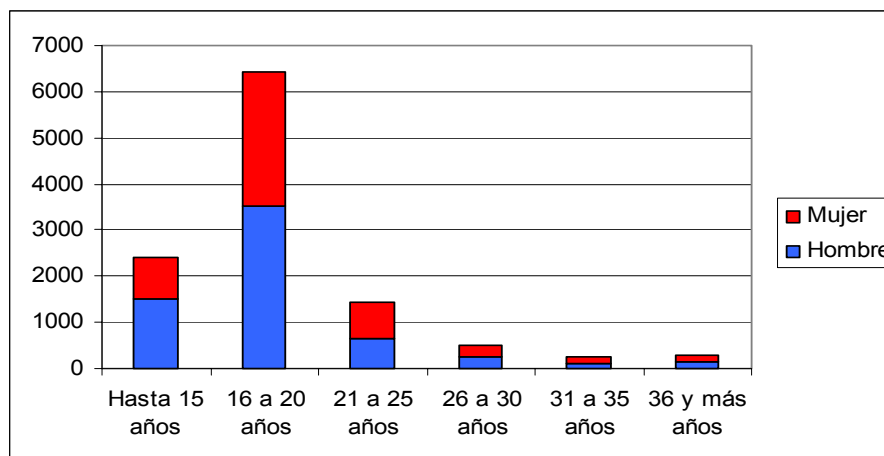


Gráfico N° 8: Número de alumnos por edad y Sexo
Fuente: Evaluación Diagnóstica Nivel II – CETP 2008

- Esos mismos datos se representa usando valores relativos. En este caso, se debe recordar que el peso de cada columna en el total de las observaciones no es la misma, pero es más fácil comparar la distribución de una de las variables en cada modalidad de la otra. En el ejemplo mostrado, puede deducirse fácilmente la tasa de masculinidad² para cada uno de los grupos etáreos.

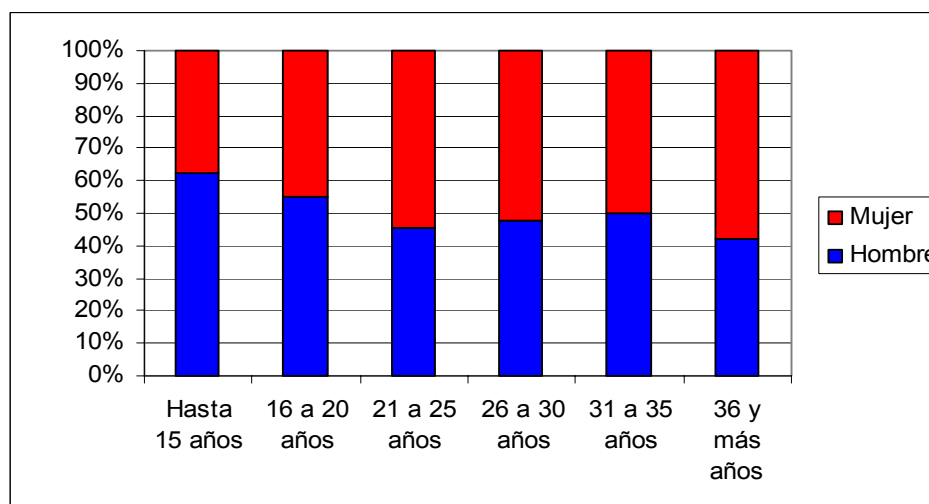


Gráfico N° 9: Porcentaje de alumnos por edad y Sexo
Fuente: Evaluación Diagnóstica Nivel II – CETP 2008

Se reitera que el gráfico debe respetar lo más fielmente posible la descripción de los datos obtenidos en la investigación, por lo cual su elección debe ser cuidadosa.

² Cantidad de varones sobre cantidad de mujeres en cierta población o sector de la población.

3. PROBABILIDAD

3.1 CONCEPTO Y AXIOMAS

Los comienzos históricos de la probabilidad se identifican generalmente con los problemas derivados de los juegos de azar.

Su formulación como una rama de la Matemática se remonta básicamente al siglo XVIII, siendo sus principales exponentes Leonhard Euler, Thomas Bayes, Jean Le Rond D’Alambert, George Louis Leclerc, Pierre Laplace y varios integrantes de la familia Bernoulli, principalmente Daniel.

Por ejemplo, Pierre Laplace, en su libro “Teoría Analítica de las Probabilidades” (1812) da la siguiente definición: “*Probabilidad* de un suceso es la razón entre el número de casos favorables y el número total de casos posibles, siempre que nada obligue a creer que alguno de estos casos debe tener lugar de preferencia a los demás, lo que hace que todos sean, para nosotros, igualmente posibles”. Si se presta atención, se puede observar que para definir “probabilidad” utiliza la expresión “igualmente posibles” o “igualmente probables”, lo que indica algunas deficiencias en la misma.

Desde comienzos del siglo XX, principalmente con los aportes de Karl Pearson, Andrei Markov y Andrei Kolmogoroff se sistematiza la Teoría de las Probabilidades, brindándole la base axiomática, que se verá a continuación.

2.1.1 Definiciones

- ✓ Se llama *espacio muestral* asociado a un experimento al conjunto de todos los resultados posibles de dicho experimento. Se lo designa habitualmente por la letra griega omega (Ω). Los experimentos que dan lugar a los espacios muestrales se llaman *experimentos aleatorios*.
Ejemplos:
 - a) Si el experimento es “tirar un dado y observar la cara que queda hacia arriba”, el espacio muestral asociado es $\Omega = \{1, 2, 3, 4, 5, 6\}$
 - b) Si el experimento es “elijo un estudiante de Educación Media y le pregunto que tipo de curso realiza”, el espacio muestral definido en los ejemplos anteriores es: $\Omega = \{\text{Ciclo Básico CES, Ciclo Básico Privado, Ciclo Básico Tecnológico UTU, Formación Profesional UTU, Formación Profesional Privado, Bachillerato Diversificado CES, Bachillerato Diversificado Privado, Bachillerato Tecnológico UTU}\}$
- ✓ Se llama *suceso o evento* a cada subconjunto de Ω . Cada uno de los resultados que componen el espacio muestral se denomina suceso elemental o caso. Se designa con A al conjunto de todos los sucesos de un experimento.
Ejemplos:
 - a) En el caso del dado, un suceso puede ser $A = \{\text{salga un número par}\}$
 - b) En el caso de los alumnos, un suceso puede ser $B = \{\text{el alumnos está cursando Ciclo Básico}\}$
- ✓ Dos sucesos se dicen incompatibles o mutuamente excluyentes si no tienen ningún elemento en común (o sea, su intersección es vacía)
Por ejemplo, en el caso del dado, son sucesos mutuamente excluyentes $A = \{\text{salga un número par}\}$ y $B = \{\text{salga 1 o 3}\}$

2.1.3 Definición axiomática

Un axioma es una propiedad que se acepta como verdadera sin demostración. En la formulación moderna de la Matemática, los axiomas son la base sobre la cual se erige toda la teoría.

Se llama Probabilidad a una función $P: A \rightarrow [0,1]$ que cumpla con las siguientes propiedades:

a1) $P(\Omega) = 1$

a2) Si A_1, A_2, A_3, \dots son sucesos incompatibles, se cumple que:

$$P\left(\bigcup A_i\right) = \sum P(A_i)$$

Estas propiedades significan que:

a1) La probabilidad de que ocurra uno cualquiera de los resultados posibles de un experimento es 1.

a2) Si dos o más sucesos son mutuamente excluyentes, o sea, no tienen elementos en común, la probabilidad de que ocurra uno cualquiera de ellos se calcula sumando las probabilidades individuales de cada uno de los mismos.

2.2 PROPIEDADES

2.2.1 Propiedades generales

A partir de los axiomas se pueden formular las siguientes propiedades:

p1) $P(\emptyset) = 0$

p2) $P(A^c) = 1 - P(A)$

p3) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

El significado de estas propiedades es el siguiente:

p1) La probabilidad de que no ocurra ninguno de los resultados posibles del experimento es 0

p2) Se llama *complemento de un suceso* a los elementos del espacio muestral que no formen parte de ese suceso. Por ejemplo: en el caso de los alumnos de Educación Media, si el suceso es "Que esté cursando Ciclo Básico", los elementos de ese suceso son que el alumno curse Ciclo Básico en Enseñanza Secundaria, en Liceos Privados o en el Consejo de Educación Técnico Profesional. El complemento de este suceso es que el alumno curse Formación Profesional o Bachillerato en cualquiera de sus opciones.

La propiedad indica que la probabilidad de que ocurra el complemento de un suceso cualquiera es 1 menos la probabilidad del suceso especificado. De otra forma: la suma de las probabilidades de dos sucesos complementarios, es 1.

p3) Si dos sucesos no son mutuamente excluyentes, la probabilidad de que ocurra uno u otro se calcula como la suma de las probabilidades individuales de cada suceso menos la probabilidad de que ocurran los resultados comunes a ambos. Un ejemplo es el siguiente: Suceso A: que el alumno curse Ciclo Básico y suceso B: que el alumno esté matriculado en UTU. Los elementos que tienen en común son los alumnos del Ciclo Básico Tecnológico.

2.2.2 Probabilidad clásica

El concepto de probabilidad de Laplace puede deducirse como una propiedad de los axiomas vistos:

Si A_1, A_2, \dots, A_n es una partición finita del espacio muestral Ω (o sea, no tienen entre ellos ningún elemento en común y la unión de todos ellos es Ω) y además se cumple que:

$P(A_1) = P(A_2) = \dots = P(A_n) = 1/n$ (o sea, son sucesos equiprobables), entonces:

$$P\left(\bigcup_{i=1}^k A_i\right) = \frac{k}{n} = \frac{N^\circ \text{casos favorables}}{N^\circ \text{casos posibles}}$$

O sea, la probabilidad de la unión (\cup) de k sucesos en n posible, se calcula como el cociente entre el número de casos "favorables" y el número de casos "posibles".

La diferencia con la definición de Laplace, es que ahora sí se tiene una definición general de probabilidad y ésta se aplica en un caso particular (aquél en el que todos los sucesos tienen igual probabilidad).

Por ejemplo: Si se quiere elegir una Escuela de UTU al azar, la probabilidad de que corresponda al departamento de Artigas se calcula como:

$$P(\text{Escuela}_{\text{ Artigas}}) = \frac{N^\circ \text{ escuelas}_{\text{ Artigas}}}{N^\circ \text{ escuelas}_{\text{ Total}}} = \frac{5}{129}$$

2.2.3 Probabilidad condicional

Se señala con el símbolo $P(A|B)$ la probabilidad de que se dé el suceso B dado que se sabe que ocurrió A.

Su definición es la siguiente:

$$P(B | A) = \frac{P(A \cap B)}{P(A)}$$

O sea, la probabilidad de que ocurra B dado que se sabe que ocurrió A se calcula como el cociente de la probabilidad de que se den los dos sucesos simultáneamente, o sea la intersección de los conjuntos, (\cap) (también llamada probabilidad conjunta) dividido la probabilidad del suceso dado.

Ejemplo

Al trabajar con los datos extraídos de la Encuesta Nacional de Hogares Ampliada, si se desea calcular la probabilidad de que un alumno elegido al azar concorra a un Centro de Ciclo Básico Tecnológico, sabiendo que es un alumno de UTU, puede calcularse de la siguiente manera:

$$P(\text{Alumno}_{\text{ CBT}} | \text{Alumno}_{\text{ UTU}}) = \frac{P(\text{alumno}_{\text{ CB de UTU}})}{P(\text{alumno}_{\text{ UTU}})}$$

Utilizando los datos extraídos de la encuesta, sabemos que:

$$P(\text{Alumno}_{\text{ CB de UTU}}) = \frac{\text{alumnos}_{\text{ CBT UTU}}}{\text{Total}_{\text{ alumno}}} = \frac{16822}{292937} = 0.074$$

$$P(\text{Alumno}_{\text{ UTU}}) = \frac{\text{Total}_{\text{ alumnos}_{\text{ UTU}}}}{\text{Total}_{\text{ alumnos}}} = \frac{48118}{292937} = 0.164$$

Sustituyendo estos valores en la fórmula inicial se encuentra:

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{0.074}{0.164} = 0.451$$

Como puede observarse, el conocimiento del suceso A tiene influencia en la probabilidad de la ocurrencia del suceso B, o sea, el saber que es un alumno de UTU aumenta la probabilidad de que el alumno concorra al Ciclo Básico Tecnológico respecto a la probabilidad del mismo suceso cuando no se tiene ningún dato.

3.2.4 Sucesos independientes

En el caso anterior, la ocurrencia de un suceso tiene repercusión en la probabilidad de la ocurrencia de otro suceso. Pero esto no siempre sucede. Se dice que dos sucesos A y B son *independientes* si la ocurrencia de uno de ellos no influye en la ocurrencia del otro. En fórmulas:

$$P(B|A) = P(B) \text{ o también: } P(A|B) = P(A)$$

En este caso se cumple que: $P(A \cap B) = P(A) \cdot P(B)$

O sea, la probabilidad de que los dos sucesos ocurran simultáneamente es igual al producto de las probabilidades de cada uno de ellos.

La idea de independencia entre variables es muy importante en Estadística, ya que si las variables no son independientes, pueden establecerse criterios de correspondencia para casos más complejos, que permitan deducir la ocurrencia de un fenómeno dado que se conoce la ocurrencia de otro u otros fenómenos relacionados con el mismo hecho.³

3.2.5 Teorema de Bayes

Si se tiene una partición finita del espacio muestral como ya se definió para el caso de los sucesos mutuamente excluyente:

B_1, B_2, \dots, B_n tales que su unión es Ω y no tienen elementos en común, y además otro suceso A del mismo espacio muestral con la única condición que su probabilidad no sea nula, se cumple que:

$$P(B_j | A) = \frac{P(B_j)P(A | B_j)}{\sum_{i=1}^n P(B_i)P(A | B_i)}$$

Esta fórmula resuelve el siguiente problema: suponiendo que un suceso A pueda producirse como consecuencia de cualquiera de los sucesos B_i y sabiendo que A se ha producido, averiguar la probabilidad de que haya sido debido a la causa B_j .

Se llama *probabilidad a priori* a la probabilidad de A dado que ocurrió B_i ($P(A|B_i)$) y probabilidad a posteriori al resultado ($P(B_j|A)$)

Esta propiedad es sumamente usada, particularmente en temas relacionados con medicina. Actualmente existe toda una escuela de Estadística que la tiene como base de su teoría.

³ Estos criterios se verán someramente al final del curso.

Introducción a la Probabilidad y Estadística

Ejemplo:

Si se trabaja con los datos de los alumnos matriculados en Educación Media, extraídos de la Encuesta Nacional de Hogares Ampliada, se definen los conjuntos B_i de la siguiente manera:

$B_1 = \{\text{alumnos del Consejo de Educación Técnico Profesional}\}$

$B_2 = \{\text{alumnos del Consejo de Educación Secundaria}\}$

$B_3 = \{\text{alumnos de Institutos Privados}\}$

$A = \{\text{alumnas de Educación Media}\}$

Si se desea calcular la probabilidad de que la persona esté inscrita en el CETP, sabiendo que es mujer, la tabla de contingencia que permite extraer los datos requeridos es la siguiente:

Organismo	Masculino	Femenino	Total
Consejo de Educación Técnico Profesional	29.010	19.108	48.118
Consejo de Educación Secundaria	91.388	113.832	205.220
Institutos Privados	19.123	20.476	39.599
TOTAL	110.511	134.308	244.819

Cuadro N° 12: Número de alumnos matriculados en Educación Media por Género según Organismo de Enseñanza
Fuente: Encuesta Nacional de Hogares Ampliadas – INE – 2006

Los cálculos necesarios para poder aplicar la fórmula del Teorema de Bayes son las siguientes:

$$P(B_1) = \text{Alumnos CETP} / \text{Total alumnos} = 0,1643$$

$$P(B_2) = \text{Alumnos CES} / \text{Total alumnos} = 0,7006$$

$$P(B_3) = \text{Alumnos Privados} / \text{Total alumnos} = 0,1352$$

$$P(A|B_1) = \text{Alumnas CETP} / \text{Total alumnos CETP} = 0,3971$$

$$P(A|B_2) = \text{Alumnas CES} / \text{Total alumnos CES} = 0,5547$$

$$P(A|B_3) = \text{Alumnas Privado} / \text{Total alumnos Privados} = 0,5171$$

Sustituyendo estos valores en la fórmula del Teorema de Bayes, se obtiene:

$$P(B_1 | A) = \frac{P(B_1)P(A | B_1)}{P(B_1)P(A | B_1) + P(B_2)P(A | B_2) + P(B_3)P(A | B_3)}$$

$$P(B_1 | A) = \frac{0.1643 * 0.3971}{0.1643 * 0.3971 + 0.7006 * 0.5547 + 0.1352 * 0.5171} = 0.1245$$

Esto quiere decir es que si se elige al azar una persona matriculada en Educación Media, la probabilidad de que esté inscripto en un curso del Consejo de Educación Técnico Profesional sabiendo que es mujer, es de **0.1245**.

3.3 DIFERENTES ENFOQUES

Existen varios enfoques para el cálculo de probabilidades. Se verán los más importantes:

- Probabilidades finitas: se basan en experimentos con un número finito de resultados que pueden resolverse de antemano mediante la definición clásica de probabilidad. En general son problemas de Análisis Combinatorio en los cuales la teoría de probabilidad sirve para darles un enunciado más atractivo. Son ejemplos típicos de estos problemas los referidos a dados y a cartas.
- Probabilidad frecuentista o experimental: la probabilidad aparece como el resultado de muchos ensayos o pruebas, sin que se pueda pensar en calcularla de antemano, ya sea por desconocer la manera de actuar de las causas que originan el fenómeno, ya sea por ser éstas demasiado numerosas o complicadas. Por ejemplo, las probabilidades halladas en los ejemplos anteriores fueron calculadas en base a datos obtenidos mediante muestras de la población.
- Probabilidad subjetiva o Bayesiana: los resultados no provienen de una estadística de casos idénticos o muy análogos, sino que se basan en opiniones emitidas por expertos en las condiciones del problema. Su formulación parte del teorema de Bayes.

3.4 DISTRIBUCIONES DE VARIABLES ALEATORIAS

Los elementos del espacio muestral son elementos abstractos y, en consecuencia, también lo son los sucesos definidos a partir de ellos. La probabilidad es una función cuyo dominio es el conjunto de los sucesos y cuyo codominio es el intervalo $[0,1]$ de los números reales. Para poder aplicar el cálculo matemático es conveniente que el dominio pertenezca también a un conjunto numérico. Para solucionar este problema suele asignarse una función del conjunto de los sucesos en el conjunto de los números (reales en general) y a estos números asignarle una probabilidad. A estas funciones se les llama *variables aleatorias*. La función que le asigna a cada valor de la variable una probabilidad, se llama *función de probabilidad* o *distribución de probabilidad* de la variable aleatoria.

Dependiendo del conjunto numérico considerado, las variables aleatorias pueden ser *discretas* (en general, son números naturales y pueden ser de recorrido finito o infinito) o *continuas* (son números reales).

3.4.1 Variables discretas

3.4.1.1 Distribución de Bernoulli

Una variable aleatoria tiene distribución de Bernoulli⁴ si puede tomar únicamente dos valores (en general representados por 0 y 1) y que son considerados tradicionalmente como “fracaso” y “éxito” respectivamente. Se caracteriza a estas variables por la probabilidad de éxito.

Por ejemplo: el caso del género de una persona es de una variable aleatoria con distribución Bernoulli.

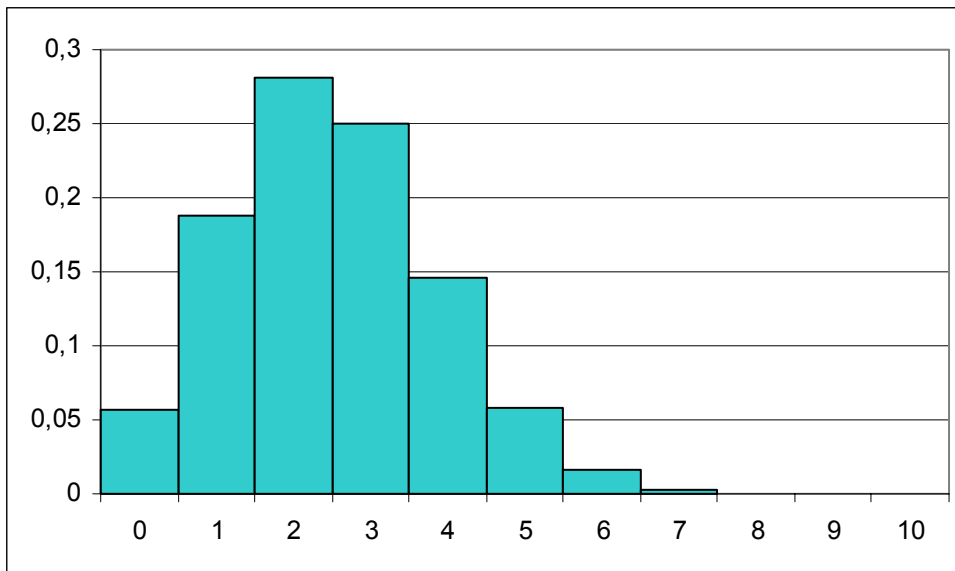
3.4.1.2 Distribución binomial

Si un experimento consta de varias pruebas independientes repetidas, teniendo cada una de ellas distribución Bernoulli, y se requiere el número de “éxitos”, la variable aleatoria asociada al experimento tiene una distribución binomial. Son necesarios dos datos: el total de pruebas y la probabilidad de éxito en cada una de ellas.

Un ejemplo de variable con distribución binomial, es el caso siguiente:

En una prueba de 10 ítems, cada uno tiene 4 opciones como respuesta, de las cuales solo una es la correcta. Si un alumno no estudió y contesta al azar, la probabilidad de “acierto” es de 0.25. Con estos datos puede calcularse la probabilidad de que un alumno que conteste al azar tenga determinado número de respuestas correctas. Estas probabilidades se muestran en el siguiente gráfico:

⁴ También son conocidas como “Variables indicatrices” y en Economía suele llamárselas “Variables Dummies”



Gráfica N° 10:
Probabilidad de número de aciertos en una prueba de 10 ítems con 4 opciones cada uno
Fuente: creación propia

3.4.1.3 Distribución de Poisson

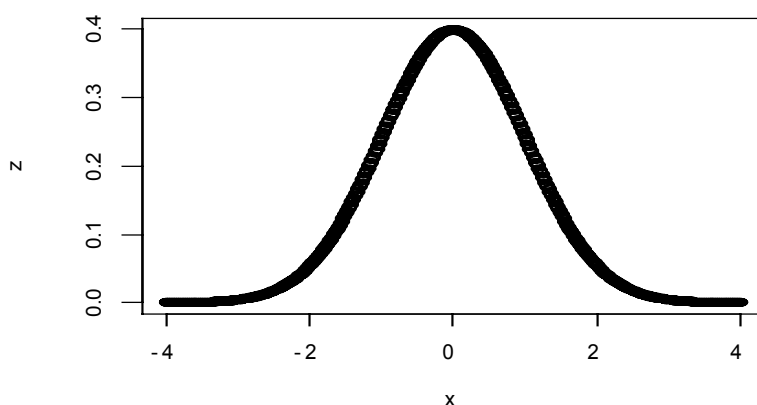
Una variable con distribución de Poisson cuenta la cantidad de sucesos ocurridos durante un período determinado de tiempo. El recorrido de esta variable es el conjunto de los números naturales.

Si, por ejemplo se considera un período fijo de un día, una variable que cuente la cantidad de alumnos que faltaron a clase tiene distribución de Poisson.

3.4.2 Variables continuas

3.4.2.1 Distribución Normal

La distribución de una variable normal se caracteriza por dos valores: la media y la variancia (ambos conceptos ya vistos). Tiene muchas aplicaciones en Estadística y la gráfica de su función de probabilidad tiene una forma conocida por “Campana de Gauss”. En una distribución normal, el 68% de los valores se encuentran en el intervalo determinado por la media más y menos el desvío estándar y el 99% de los valores están en el intervalo determinado por la media más y menos tres veces el desvío.



Gráfica N° 11: Distribución de una variable Normal con media igual a cero y desvío estándar igual a uno (Campana de Gauss)

Existe una propiedad muy importante, llamada “*Teorema Central del Límite*” que dice que si se tiene la media de un número muy importante de variables aleatorias independientes y con la misma distribución, la distribución de estas medias tiende a ser normal. (Cuidado: las variables originales siguen teniendo la distribución original, lo que tiende a ser normal es la nueva variable aleatoria determinada por las medias).

3.4.2.2 Distribución Exponencial

La distribución exponencial se utiliza básicamente para medir tiempos o distancias entre dos sucesos del mismo tipo.

Existe una relación muy importante entre la distribución Exponencial y la Poisson: si una variable tiene distribución de Poisson, el tiempo que separa la ocurrencia de dos eventos es una variable con distribución Exponencial.

3.4.2.3 Otras distribuciones

Algunas de las distribuciones continuas más usadas en Estadística se refieren a variables que son necesarias para la solución de problemas, tales como la independencia de variables o verificación de hipótesis de trabajo. Entre ellas se destacan la distribución χ^2 (chi cuadrado) y la distribución **t de Student**

En general, para comprobar que una variable tiene determinada distribución, deben realizarse una serie de pruebas, llamadas *pruebas de bondad de ajuste*.

4 INFERENCIA ESTADÍSTICA

4.1 INTRODUCCIÓN

En muchas oportunidades es imposible acceder a todos los datos de una población determinada (ya sea por costos materiales o de tiempo o porque al obtener el dato el elemento se destruye). En estos casos debe trabajarse con una muestra de la población en cuestión. Cuando los datos obtenidos en esa muestra se extienden a la población de origen, se realiza una *Inferencia Estadística*.

Los datos obtenidos en la muestra son variables aleatorias. Toda función de esas variables que no dependan de valores desconocidos se denomina *estadístico*. Los estadísticos pueden ser valores tales como la media, un total, un porcentaje, etc. Cuando un estadístico se utiliza para estimar un valor poblacional se llama *estimador* de dicho valor. Al valor desconocido suele llamársele *parámetro*.

Las estimaciones de un valor desconocido de la población se pueden realizar de dos maneras: estimación puntual y estimación por intervalos.

4.2 ESTIMACIÓN PUNTUAL

El método de la estimación puntual consiste en elegir la función de datos de la muestra cuyo valor, con cierta probabilidad, pueda tomarse como valor de la población.

Las características favorables de un estimador son:

- Insesgado – Un estimador se dice insesgado si cuando se toman muchas muestras en lugar de una sola, el promedio de los valores del estimador es el valor poblacional. Por ejemplo, si lo que se quiere estimar es la media de determinada variable de la población, se puede considerar que la media de la muestra es un estimador insesgado.
- Eficiente (o de mínima varianza) – Si dos estimadores del mismo valor poblacional son insesgado, para elegir cual es mejor pueden compararse las varianzas de ambos y elegir el que tenga menor valor.
- Estimador de máxima verosimilitud – Un estimador es máximo verosímil si se elige el valor del parámetro para el cual es máxima la probabilidad de haber sacado la muestra obtenida.

4.3 ESTIMACION POR INTERVALO

En algunas ocasiones es conveniente dar, además de un valor puntual del valor poblacional, un intervalo, o sea, un par de valores dentro de los cuales la probabilidad de que se encuentre el parámetro es conocida. Por ejemplo, si los extremos del intervalo son a y b , se dice que la probabilidad de que el valor buscado esté entre a y b es de 95%. En este caso se dice que $[a, b]$ es un *intervalo de confianza* del 95% para el valor buscado. Esto quiere decir que si en lugar de tomarse una muestra, se toman 100, un 95% de las veces el estimador se encuentra entre a y b .

En general, una forma de obtener el intervalo de confianza, es hallar un valor puntual del mismo. El intervalo final se obtiene sumando y restando cierto número al estimador puntual. En estos casos se usa cierta distribución (como la Normal o la t de Student) y la varianza de la muestra para obtener este número. Para que la estimación por intervalos sea útil, el rango del intervalo no debe ser demasiado grande.

4.4 PRUEBAS DE HIPÓTESIS

Una hipótesis es una afirmación respecto a un parámetro poblacional. La finalidad de una prueba de hipótesis es decidir, basándose en una muestra de la población, cual de dos afirmaciones complementarias es verdadera. Las dos hipótesis complementarias se llaman hipótesis nula (que en general se indica H_0) e hipótesis alternativa (H_1).

Una forma muy usual de plantear la prueba de hipótesis es:

H_0 : el valor del parámetro es θ

H_1 : el valor del parámetro es distinto de θ

En este caso se dice que es una prueba “a dos colas”, ya que no importa si el valor desconocido es mayor o menor que θ , sino simplemente si es distinto.

Otra posibilidad es:

H_0 : el valor de parámetro es θ

H_1 : el valor del parámetro es mayor que θ

En este caso, debemos verificar si el valor que buscamos es igual a θ o mayor que él, por lo que se habla de una prueba “a una cola”.

En todos los casos existen distribuciones para el valor del parámetro. Se debe comparar los resultados obtenidos por el test con un valor dado de antemano, llamado *valor crítico*. A partir de este punto se establecen dos regiones llamadas Región de Aceptación y Región de Rechazo. Si el valor del test cae dentro de la región de aceptación, se acepta la hipótesis nula, de lo contrario, se rechaza.

Como en estos casos se trabaja con una muestra, se puede cometer algún error. Si se rechaza H_0 cuando en realidad es verdadera, se habla de *error tipo I*. Por el contrario, si se acepta H_0 cuando es falsa, se dice que se cometió un *error tipo II*.

4.5 TEST DE INDEPENDENCIA

4.5.1 Variables cualitativas

En algunas ocasiones interesa saber si dos variables son o no independientes.⁵ Se trabaja en estos casos con una prueba de hipótesis especial, en la cual la H_0 es que las variables son independientes y la H_1 que no lo son. Se denominan test de independencias de Pearson.

En este test se trabaja con tablas de doble entrada, en las cuales se coloca las distintas modalidades de una variable como filas y las modalidades de la otra variable como columnas. La condición principal es que no haya datos faltantes, ya que se necesitan los diferentes cruzamientos de las mismas (o sea, cuantos elementos de la muestra comparten cada combinación de las diferentes modalidades de ambas variables). También debe elegirse un valor de significación del test (o sea, cual es el mayor error que se permite).

Se verá como trabaja este método con un ejemplo:

La pregunta es si las diferencias se deben a la muestra o si es que las variables no son efectivamente independientes. El matemático Karl Pearson demostró en 1901 que si se efectúan las siguientes operaciones:

⁵ Se recuerda que dos variables son independientes si la información que se posee sobre una de ellas no influye en el resultado de la otra.

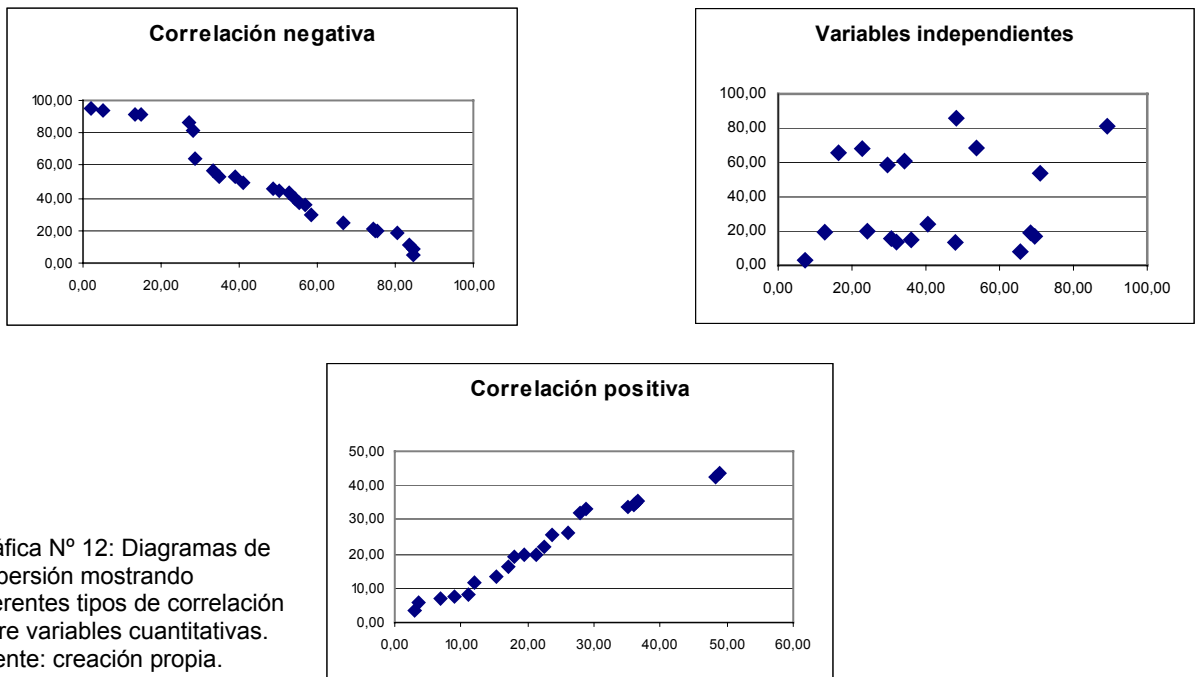
$$\sum \frac{(\text{valor. esperado} - \text{valor. observado})^2}{\text{valor. esperado}}$$
 la variable aleatoria obtenida se distribuye χ^2 cuyos grados de libertad se obtienen como $(N^\circ \text{ filas} - 1)(N^\circ \text{ columnas} - 1)$

3.5.2 Variables cuantitativas

Si dos variables aleatorias referidas a datos sobre los mismos elementos de la población son cuantitativas, puede establecerse que grado de relación tienen mediante el llamado *coeficiente de correlación*. Es un número entre -1 y 1 y sus valores se pueden interpretar de la siguiente forma:

- Si el coeficiente de correlación es 0 , las variables son independientes.
- Si el coeficiente de correlación es 1 , las variables están correlacionadas positivamente (o sea, a mayores valores de una de ellas corresponden mayores valores de la otra).
- Si el coeficiente de correlación es -1 , las variables están correlacionadas negativamente (o sea, a mayores valores de una de ellas corresponden menores valores de la otra)
- Para los valores intermedios, no hay correlación o independencia perfecta, pero en general, si el valor es cercano a 1 o a -1 , se acostumbra a decir que las variables tienen correlaciones altas. No existe un valor límite para separar estos conceptos, pero puede considerarse que valores mayores a $0,7$ ya indican altas correlaciones y valores menores a $0,3$ indicarían la ausencia de correlación (o cuasi independencia).

Gráficamente, puede usarse el llamado *diagrama de dispersión* o dispersograma, en el cual cada variable se representa en uno de los ejes y cada punto queda determinado por los valores de cada variable para una observación.



Gráfica N° 12: Diagramas de dispersión mostrando diferentes tipos de correlación entre variables cuantitativas. Fuente: creación propia.

4 MUESTREO ALEATORIO

4.1 CONCEPTOS BÁSICOS

4.1.1 Población

Una *población finita* (o simplemente población) es un conjunto finito de elementos. Esto implica que puede determinarse sin ambigüedad si un elemento pertenece o no al conjunto. Denominaremos **U** a nuestra *población objetivo*, o sea la población finita sobre la cual deseamos obtener alguna información.

Al número de elementos de la población lo anotaremos **N** y, en general, lo supondremos conocido. Así mismo, supondremos que cada elemento de la población es identificable y que se le puede otorgar un número natural del 1 al **N**.

Ejemplos:

- Docentes de la Administración Nacional de Educación Pública
- Alumnos matriculados en el Ciclo Básico de Educación Media

4.1.2 – Encuesta por muestreo

Teóricamente una *muestra* es un subconjunto de la población. La esencia de una *encuesta por muestreo* consiste en la selección de una muestra con el objetivo de establecer conclusiones sobre toda la población basándose en la información de la parte observada. Si la muestra coincide con toda la población objetivo, se denomina *censo*.

¿Por qué usar una encuesta por muestreo y no un censo?

Existen diferentes razones, entre las que podemos mencionar:

- Naturaleza destructiva de ciertas pruebas (por ejemplo en los casos de control de calidad o exámenes de laboratorio de un paciente).
- Imposibilidad física de revisar todos los elementos de la población.
- Costos prohibitivos de estudiar a todos los integrantes de una población.
- Tiempo necesario para entrevistar a todos los elementos de la población.
- Lo adecuado de los resultados de la muestra: se puede inferir los valores poblacionales de interés de acuerdo a los resultados obtenidos con la muestra.

4.1.3 – Marco muestral

Se define como el conjunto de conjunto de unidades, procedimientos y mecanismos que identifican, distinguen y permiten acceder a la población objetivo.

Físicamente la muestra se extrae de este listado, que por diversas razones (desactualizaciones, duplicaciones, errores, etc.) puede no coincidir con la población objetivo.

4.1.4 – Muestreo probabilístico

Definimos muestro *probabilístico* como una selección de muestra que cumpla:

- i) El conjunto de todas las muestras posibles es conocido.
- ii) Cada muestra tiene una probabilidad conocida de selección. El procedimiento de selección asigna a cada elemento de la población una probabilidad no nula de ser incluido en la muestra.
- iii) Se selecciona una muestra por un mecanismo aleatorio bajo el cual cada muestra posible tiene exactamente la probabilidad $p(s)$ de ser extraída.

A una muestra obtenida bajo las condiciones anteriores se le denomina *muestra aleatoria* o *muestra probabilística*.

4.2 DIFERENTES DISEÑOS DE MUESTREO

4.2.1 Muestreo Aleatorio Simple

De una población **U** de **N** elementos, se extraen **n** de manera independiente y sin reponer (o sea, el elemento que ya ha sido extraído no influye en la extracción de los siguientes y tampoco puede volverse a elegir) Para efectuar esta elección se pueden utilizar tablas de números aleatorios, aunque los distintos software estadísticos disponen de rutinas que permiten este tipo de muestreo.

La extracción debe realizarse a partir de un marco muestral de lista: cada una de las unidades están identificadas con un número (del 1 al **N**) y se sortean **n** elementos.

La probabilidad de inclusión de un elemento **k** es: $P(k \in s) = n / N$ por lo que todos las unidades del marco muestral tienen la misma probabilidad de ser extraídas.

4.2.2 Muestreo sistemático

En algunas ocasiones, los elementos de la población están ordenados según un criterio determinado (alfabético, por fecha, por monto, etc.) En estos casos, no siempre es recomendable efectuar un muestreo aleatorio simple.

Un diseño de muestreo aplicable en esta situación es el siguiente:

Se elige un número natural **a** (fijo) llamado *intervalo de muestreo*. Se elige aleatoriamente un número natural **r** entre 1 y **a** (llamado *arranque aleatorio*) La muestra se forma por todos los elementos de la población cuyo número de identificación coincida con $r + (h - 1) \cdot a$ con $h = 1, 2, \dots, n$ siendo **n** el cociente por defecto de N / a .

En total hay **a** muestras posibles. De otra manera, el primer integrante de la muestra será el que ocupe el lugar **r**, luego el que ocupe el lugar **r+a**, luego **r+2a**, etc.

¿Cuándo usar Muestreo Sistemático y cuándo Muestreo Aleatorio Simple?

Se debe tener cuidado cuando la distribución de los elementos en la población presenta ciclos en los valores de la variable de interés, ya que en el muestreo sistemático se extraerán valores semejantes, lo que se transfiere a una sub o super valoración del valor verdadero. En estos casos es más conveniente usar el muestreo aleatorio simple.

Por lo tanto:

- Si la distribución de la variable en la población ordenada es aleatoria, los dos diseños de muestreo tienen la misma performance. A veces redundante en una economía de recursos el uso del Muestreo Sistemático.
- Si la variable presenta ciclos o estacionalidades, es más eficiente el Muestreo Aleatorio Simple.
- Si la distribución de la variable es creciente o decreciente (por ejemplo, montos de deudores o acreedores) es más eficiente el Muestreo Sistemático.

4.2.3 Muestreo estratificado

El muestreo estratificado consiste en dividir la población en H grupos o subpoblaciones llamados estratos o unidades primarias y tomar una muestra independiente de manera aleatoria en cada una de ellas. La estratificación puede realizarse utilizando diferentes variables, dependiendo del objetivo planteado al realizar la muestra, como por ejemplo: por departamento o región geográfica, por edad, por género, por nivel socio económico, por tipo de curso que realice el estudiante, etc. Es una herramienta poderosa y flexible, muy comúnmente usada en la práctica. Algunas de las razones principales son:

- En una población heterogénea se puede efectuar una buena partición por medio de una variable auxiliar.
- Se permite distintas precisiones por estratos.
- Aspectos tales como la no-respuesta, facilidad de locación y disponibilidad de información auxiliar puede diferir entre estratos y este tipo de muestreo permite elegir el diseño que mejor se adapte en cada caso.
- Razones administrativas o de costos en la organización de la encuesta pueden hacer que resulta prácticamente imposible la utilización de un Muestreo Aleatorio Simple.

Existen criterios matemáticos para determinar el número de estratos que optimice la variancia de la variable de interés, así como de los puntos de corte entre estos estratos. Para ello se debe utilizar como insumo la distribución de la variable obtenida de un censo reciente o de otra encuesta. También se puede utilizar una variable auxiliar que esté altamente correlacionada con la variable objetivo.

4.2.4 Muestreo por conglomerados y en varias etapas

Hasta ahora hemos visto diseños de muestreo que asumen que se puede realizar un muestreo directo de elementos. Sin embargo, en las encuestas de mediana y gran escala esto no siempre es posible ya sea porque no se dispone de un marco que identifique a todos los elementos y el costo de crear uno es demasiado elevado o los elementos de la población están muy dispersos en un área geográfica muy extensa por lo que el muestro directo de elementos lleva a costos de relevamiento excesivamente altos.

Los diseños de muestreo en dos etapas y multietapa no requieren realizar muestreo directo de elementos, ya que una primera etapa se muestrean grupos (o clusters) de elementos, o sea, son aplicables cuando se poseen marcos agrupados. Por ejemplo, se conocen los grupos con los que cuenta un establecimiento de enseñanza, pero no los nombres de los alumnos inscriptos en cada uno de ellos.

La diferencia entre Muestreo Estratificado y Muestreo en varias etapas es que en el primero extraemos una muestra de todos los estratos, mientras que en el segundo, primero se hace una muestra de las Unidades Primarias de muestreo y luego, en las seleccionadas exclusivamente, se realiza el muestreo de unidades (o se continua en otras etapas). Es muy común en la práctica que en la primera etapa del muestreo se realice una estratificación y luego se extraigan muestras de cada uno de esos estratos.

5 TÉCNICAS MULTIVARIADAS

5.1 MODELOS DE REGRESIÓN LINEAL

En ciertas ocasiones es necesario conocer el comportamiento de una variable en su relación con otras. En estos casos puede determinarse qué influencias tienen las variables dadas (llamadas explicativas o independientes) en la variable de estudio (llamada variable de respuesta o dependiente). Una vez establecida esta relación, se puede predecir el comportamiento de la variable dependiente si se conocen los valores de las variables explicativas para un caso puntual. Si todas las variables intervinientes son cuantitativas (siendo continua la variable de respuesta y continuas o indicatrices las variables explicativas) puede estudiarse bajo un modelo de Regresión Lineal.

Estos modelos tienen como formulación matemática una ecuación del tipo:

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki} + \varepsilon_i$$

En donde y_i es la variable de respuesta para el caso i

x_{ji} son las diferentes variables explicativas (en total son k)

ε_i es el error cometido al explicar la variable y mediante las variables x_j (llamado residuo)

β_j son los parámetros de regresión, a determinar por el modelo. Se interpretan como la variación de la variable de respuesta por cada unidad que cambia la variable x_j , quedando las demás constantes.

No siempre es posible aplicar modelos de regresión lineal. Deben darse ciertas condiciones que permiten que el modelo se ajuste a los datos. Por ejemplo, debe estudiarse la distribución de los residuos, que deben tener media igual a 0 y la misma variancia. En general se busca que los residuos tengan distribución normal, ya que ello facilita el resto de los cálculos. También deben estudiarse si existen observaciones "outliers", o sea, que se separan del resto de los datos, que pueden influir en el correcto ajuste del modelo.

5.2 MODELOS DE REGRESIÓN LOGÍSTICA

En algunos modelos lineales, la variable de respuesta tiene solamente dos valores posibles (que pueden considerarse como éxito y fracaso en líneas generales). En estos casos se busca que las variables independientes expliquen la probabilidad de éxito de la variable de respuesta. No puede entonces aplicarse los modelos lineales convencionales, ya que el valor obtenido no va a estar en general entre 0 y 1, lo que es imprescindible para que sea una probabilidad. Se puede usar entonces, un modelo de regresión logística. Una ecuación parecida a la vista para el modelo lineal se utiliza para obtener el cociente de probabilidades entre éxito y fracaso. En la regresión logística, las variables de respuesta pueden ser tanto cuantitativas como cualitativas. El modelo también permite predecir el comportamiento de nuevos datos, dados los valores de las variables independientes.

Si la variable de respuesta tiene más de dos categorías posibles, puede instrumentarse un modelo de regresión logística llamada multinomial. Debe tenerse cuidado con la distribución de la variable de respuesta, ya que si una de las categorías tiene muy poca frecuencia, el modelo presenta problemas de ajuste.

5.3 ANÁLISIS FACTORIAL

El Análisis Factorial es una técnica multivariada que parte de una tabla en la que se cruzan individuos (observaciones) y variables. Se pueden construir dos nubes: la de individuos (filas) y la de variables (columnas). Cada punto de la nube de individuos tiene tantas coordenadas como el número de variables. Por ejemplo, si nuestra matriz tiene I individuos y J variables, cada punto de la nube de individuos tendrá J coordenadas. A su vez, cada punto de la nube de columnas posee un número de coordenadas igual al número de individuos (en nuestro ejemplo, tendrá I coordenadas).

Los objetivos del Análisis Factorial son:

- Eliminar información redundante creando nuevas variables que no estén correlacionadas y expliquen la información que surge de la nube.
- Simplificar (reducir dimensiones) considerando la menor pérdida posible de información. Esto ocurre si la distancia entre los puntos originales y su proyección sobre los espacios creados por las nuevas variables son mínimas.
- Diferenciar de mejor manera los sujetos analizados. Para ello se construyen ejes tales que expliquen la mayor parte de la inercia o varianza de la nube.
- En relación con los individuos se trata de evaluar su semejanza. Dos individuos se asemejan más cuando más próximos sean sus valores en el conjunto de las variables.
- Con respecto a las variables, lo que se trata de evaluar es su relación.

Algunas de las técnicas particulares del Análisis Factorial son:

5.3.1 Análisis de Componentes Principales (ACP)

Se aplica a tablas que cruzan individuos (filas) y variables **cuantitativas** (columnas). En ACP la relación entre dos variables se mide por el coeficiente de correlación lineal.

El objetivo principal del ACP es proporcionar representaciones planas (o, a lo sumo, espaciales) aproximadas de la nube de individuos. Para ello se buscará una sucesión de direcciones llamadas ejes factoriales, que sean combinaciones lineales de las variables originales, perpendiculares entre sí, y que deformen lo menos posible la nube original. Con esto estamos diciendo que, si dos puntos están originariamente cercanos, en la nueva representación deben aparecer como cercanos, pero que también si estaban muy alejados originariamente, también deben aparecer alejados en la nueva representación.

5.3.2 Análisis de Correspondencia Simple (ACS)

Se trabaja sobre tablas de contingencia. Estas tablas cruzan dos variables cualitativas definidas sobre una población de n individuos.

El objetivo es medir la correspondencia o asociación entre dos variables cualitativas. También aquí se trata de construir ejes factoriales que resuman la información. Estos ejes ponen en evidencia aquellos perfiles que representan una porción más importantes de la población. Se debe tener cuidado con las modalidades (categorías) “raras” (muy pocos individuos) porque afectan la elección de los ejes. En estos casos pueden recategorizarse la variable que presente este problema.

5.3.3 Análisis de Correspondencia Múltiple (ACM)

Es una técnica factorial que trabaja sobre tablas de individuos – variables, pero estas tablas son lógicas (o sea, tienen un 1 si es individuo posee esa modalidad y 0 si no la posee). La suma de las filas es siempre igual al número de variables. La diferencia con el ACS es que puedo trabajar con cualquier número de variables y no se restringe a dos.

El objetivo del ACM respecto a los individuos es caracterizarlos. Dos individuos son más próximos cuanto tengan mayor número de modalidades en común. En cuanto a las variables, existen dos puntos de vista: estudiar la relación entre ellas mediante una tabla de datos que crucen modalidades, y sintetizar el conjunto de variables en un pequeño grupo representativo. Se dice que dos modalidades están más cercanas cuando estén presentes o ausentes simultáneamente en un gran número de individuos. En la representación gráfica de las modalidades en el plano factorial, se puede observar la cercanía o lejanía de las distintas modalidades.

5.4 ANÁLISIS DE CLUSTERS

El objetivo del Análisis de clusters (o conglomerados) es formar grupos de acuerdo a características que puedan ser de interés.

Se trabaja exclusivamente con variables cuantitativas. La formación de los grupos depende de las variables que se considere.

Los métodos para la formación de los conglomerados pueden ser jerárquicos (los grupos obtenidos a cierto nivel de distancia comprenden grupos obtenidos a un nivel inferior) o no jerárquicos (los grupos ya formados no necesariamente son mantenidos al formarse nuevos grupos). Los modelos jerárquicos son más usados por su facilidad computacional.

Generalmente se parte de un Análisis Factorial para realizar el análisis de clusters.

5.5 ANÁLISIS DISCRIMINANTE

Dado un conjunto de variables relacionadas con un tema de interés y dado un cierto número de grupos en los que se divide la población, se quiere construir un modelo que permita asignar con el menor error posible cada individuo a un grupo determinado.

Puede partirse de un Análisis Factorial (puesto que ya tenemos resumidas las variables). La partición en grupos de la población puede deberse a un Análisis de Clusters ya realizado o a grupos naturales dentro de la población de interés.

Cuando se trabaja con Análisis Discriminante se tratan los datos como si fueran la población objetivo.

El Análisis Discriminante permite hacer predicciones: si se ingresan los datos de un individuo que no participó en el análisis original, el método permite clasificarlo en uno de los grupos, indicando la probabilidad de pertenencia (en cierto sentido, el error que puede cometerse en esa clasificación).

BIBLIOGRAFÍA

- Casella, G. Berger, R.L. – Statistical Inference
- Cabaña, Enrique – Apuntes para el curso de Probabilidad I – Licenciatura en Estadística – FCEA
- Escoffier Brigitte, Pages Jerome. Análisis Factoriales Simples y Múltiples. Universidad del País Vasco. Bilbao. 1992
- Hosmer, David W.; Lemeshow, Stanley – Applied Logistic Regression – 2nd edition (2000)
- Rencher, Alvin C. – Linear Model in Statistics
- Santaló, Luis A. – Probabilidad e inferencia Estadística – Monografía N° 11, Serie Matemática – Secretaría General de la OEA
- Särndal, Carl-Erik; Swensson, Berngt; Wretman, Jan –Springer Series in Statistics