**(PORTADA)**

Manual de Métodos Cualitativos para las Ciencias Sociales

Universidad Miguel de Cervantes

Nicolás Barrientos Oradini y Sebastián Castillo Ramos

(INTERIOR DE PORTADA)

De los Autores

Nicolás Barrientos Oradini es Administrador Público, Licenciado en Ciencias de la Administración de la Facultad de Administración y Economía de la Universidad de Santiago de Chile, Magíster en Gobierno y Gerencia Pública de la Universidad de Chile, Magíster en Ciencias de la Educación con Mención en Pedagogía Universitaria de la Universidad Miguel de Cervantes, con estudios de Doctorado en Educación. Ha desempeñado cargos académicos y de gestión en universidades chilenas. Actualmente es Director de Estudios de la Universidad Miguel de Cervantes y Docente de la Escuela de Economía y Negocios. Se ha especializado en Políticas Públicas, Administración Estratégica, Educación Superior y Técnicas de Investigación Universitaria.
nbarrientos@umcervantes.cl

Sebastián Castillo Ramos es Ingeniero Comercial Mención Economía, Licenciado en Ciencias Económicas de la Universidad Alberto Hurtado, Magíster en Economía de la Universidad Alberto Hurtado y Master of Arts in Economics de la Universidad de Georgetown. Actualmente es Director de la Escuela de Recursos Humanos, Docente de la Escuela de Economía y Negocios y Ciencias Políticas y Administración Pública de la Universidad Miguel de Cervantes, además de ser Investigador Asociado de la Dirección de Estudios de la misma Universidad. Se ha especializado en Microeconometría, Economía de la Educación, Políticas Públicas y Economía del Desarrollo.
scastillo@umcervantes.cl

Universidad Miguel de Cervantes
Dirección de Estudios
Enrique Mac Iver 370, Santiago
Fono: 29273400
estudios@umcervantes.cl
www.umcervantes.cl

# Tabla de contenido

**CAPÍTULO TRES**
**Probabilidades**



**CAPÍTULO CUATRO**
**Distribuciones**

# CAPÍTULO UNO

## Uso de los Métodos Cuantitativos y la Estadística

El siguiente capítulo busca describir los usos que los métodos cuantitativos y la estadística tienen en el desarrollo de las ciencias sociales.

Concluido el capítulo se espera que usted logre los siguientes objetivos:

- Comprender el sentido de los métodos cuantitativos y su uso en las ciencias sociales.
- Comprender la lógica analítica propia de la estadística descriptiva e inferencial.
- Poner de relieve la utilidad y valor de los conocimientos que aporta para su formación académica y profesional.
- Hacerle ver que es capaz de alcanzar los objetivos y competencias de las diversas asignaturas estadísticas.

## 1.1- Introducción

Para muchos alumnos, la asignatura de Estadística resulta ser una de las más difíciles. En ocasiones, la dificultad de la misma es intrínseca, es decir, se encuentra ligada a sus objetivos, contenidos y niveles de exigencia; sin embargo, en otras, deriva de una previa actitud de “respeto” hacia los números -material con el que se trabaja en la asignatura- probablemente relacionada con carencias en la formación previa, y a la falta de sentido que muchos estudiantes le atribuyen en cuanto a las aportaciones para su formación académica y profesional. Pues bien: lo deseable es iniciar el estudio de la asignatura con una actitud semejante, en lugar de llegar a ella después de la experiencia más o menos prolongada de estudio por pura obligación. Ese es el objetivo primario de este texto.

Por ello, el enfoque del Manual no se centra en el cultivo de destrezas de cálculo y utilización de fórmulas, sino en la comprensión de sus procedimientos, procesos y aportaciones, obviando cuanto sea posible el estudio teórico. Y digo “cuanto sea posible” porque, como alguien ha dicho, la mejor práctica es una buena teoría.

## 1.2- La Estadística Descriptiva

La estadística descriptiva es la rama de las Matemáticas que recolecta, presenta y caracteriza un conjunto de datos (por ejemplo, edad de una población, altura de los estudiantes de una escuela, temperatura en los meses de verano, etc.) con el fin de describir apropiadamente las diversas características de ese conjunto.

Al conjunto de los distintos valores numéricos que adopta un carácter cuantitativo se llama variable estadística

Las variables pueden ser de dos tipos:

Variables cualitativas o categóricas : no se pueden medir numéricamente (por ejemplo: nacionalidad, color de la piel, sexo).

Variables cuantitativas: tienen valor numérico (edad, precio de un producto, ingresos anuales).

Las variables también se pueden clasificar en:

Variables unidimensionales: sólo recogen información sobre una característica (por ejemplo: edad de los alumnos de un curso).

Variables bidimensionales: recogen información sobre dos características de la población (por ejemplo: edad y altura de los alumnos de un curso)

Variables pluridimensionales: recogen información sobre tres o más características (por ejemplo: edad, altura y peso de los alumnos de una clase).

Por su parte, las variables cuantitativas se pueden clasificar en discretas y continuas:

Discretas: sólo pueden tomar valores enteros (1, 2, 8, -4, etc.). Por ejemplo: número de hermanos (puede ser 1, 2, 3...., etc., pero, por ejemplo, nunca podrá ser 3.45).

Continuas: pueden tomar cualquier valor real dentro de un intervalo. Por ejemplo, la velocidad de un vehículo puede ser 90.4 km/h, 94.57 km/h, etc.

Cuando se estudia el comportamiento de una variable hay que distinguir los siguientes conceptos:

Individuo: cualquier elemento que porte información sobre el fenómeno que se estudia. Así, si estudiamos la altura de los niños de una clase, cada alumno es un individuo; si se estudia el precio de la vivienda, cada vivienda es un individuo.

Población: conjunto de todos los individuos (personas, objetos, animales, etc.) que porten información sobre el fenómeno que se estudia. Por ejemplo, si se estudia el precio de la vivienda en una ciudad, la población será el total de las viviendas de dicha ciudad.

Muestra: subconjunto que seleccionado de una población. Por ejemplo, si se estudia el precio de la vivienda de una ciudad, lo normal será no recoger información sobre todas las viviendas de la ciudad, sino que se suele seleccionar un subgrupo (muestra) que se entienda que es suficientemente representativo.

Las variables aleatorias son variables que son seleccionadas al azar o por procesos aleatorios.

### 1.2.1- Aplicaciones y Uso de Datos

Los datos son medidas y/o números recopilados a partir de la observación. Los datos pueden concebirse como información numérica necesaria para ayudar a tomar una decisión con más bases en una situación particular.

Existen muchos métodos mediante los cuales se pueden obtener datos necesarios. Primero, se puede buscar datos ya publicados por otras fuentes. Segundo, se puede diseñar un experimento. En tercer lugar, se puede conducir un estudio. Cuarto, se pueden hacer observaciones del comportamiento, actitudes u opiniones de los individuos en los que se está interesado.

Los datos se pueden clasificar en:

• Datos discretos. Son respuestas numéricas que surgen de un proceso de conteo.

• Datos continuos. Son respuestas numéricas que surgen de un proceso de medición.

### 1.2.2 - Escalas de Medición

Medir en el campo de las ciencias exactas es comparar una magnitud con otra, tomada de manera arbitraria como referencia, denominada patrón y expresar cuántas veces la contiene. En el campo de las ciencias sociales medir es “el proceso de vincular conceptos abstractos con indicadores empíricos”. Al resultado de medir se le llama medida.

La medición de las variables puede realizarse por medio de cuatro escalas de medición: la nominal, ordinal, de intervalo y de razón. Se utilizan para ayudar en la clasificación de las variables, el diseño de las preguntas para medir variables, e incluso indican el tipo de análisis estadístico apropiado para el tratamiento de los datos.

Una característica esencial de la medición es la dependencia que tiene de la posibilidad de variación. La validez y la confiabilidad de la medición de una variable depende de las decisiones que se tomen para operarla y lograr una adecuada comprensión del concepto evitando imprecisiones y ambigüedades, en caso contrario, la variable corre el riesgo inherente de ser invalidada debido a que no produce información confiable.

a) Medición Nominal.

En este nivel de medición se establecen categorías distintivas que no implican un orden específico. Por ejemplo, si la unidad de análisis es un grupo de personas, para clasificarlas se puede establecer la categoría sexo con dos niveles, masculino (M) y femenino (F), los encuestados sólo tienen que señalar su género, no se requiere de un orden real.

Así, se pueden asignar números a estas categorías para su identificación: 1=M, 2=F o bien, se pueden invertir los números sin que afecte la medición: 1=F y 2=M. En resumen en la escala nominal se asignan números a eventos con el propósito de identificarlos.

b) Medición Ordinal.

Se establecen categorías con dos o más niveles que implican un orden inherente entre si. La escala de medición ordinal es cuantitativa porque permite ordenar a los eventos en función de la mayor o menor posesión de un atributo o característica. Por ejemplo, en las instituciones escolares de nivel básico suelen formar por estatura a los estudiantes, se desarrolla un orden cuantitativo pero no suministra medidas de los sujetos. Estas escalas admiten la asignación de números en función de un orden prescrito. Las formas más comunes de variables ordinales son ítems (reactivos) actitudinales estableciendo una serie de niveles que expresan una actitud de acuerdo o desacuerdo con respecto a algún referente. Por ejemplo, ante el reactivo: CODELCO debe privatizarse, el respondiente puede marcar su respuesta de acuerdo a las siguientes alternativas:

- Totalmente de acuerdo
- De acuerdo
- Indiferente
- En desacuerdo
- Totalmente en desacuerdo

Las anteriores alternativas de respuesta pueden codificarse con números que van del uno al cinco que sugieren un orden preestablecido pero no implican una distancia entre un número y otro.

c) Medición de Intervalo.

La medición de intervalo posee las características de la medición nominal y ordinal. Establece la distancia entre una medida y otra. La escala de intervalo se aplica a variables continuas pero carece de un punto cero absoluto. El ejemplo más representativo de este tipo de medición es un termómetro, cuando registra cero grados centígrados de temperatura indica el nivel de congelación del agua y cuando registra 100 grados

centígrados indica el nivel de ebullición, el punto cero es arbitrario no real, lo que significa que en este punto no hay ausencia de temperatura.

d) Medición de Razón.

Una escala de medición de razón incluye las características de los tres anteriores niveles de medición (nominal, ordinal e intervalo). Determina la distancia exacta entre los intervalos de una categoría.

Adicionalmente tiene un punto cero absoluto, es decir, en el punto cero no existe la característica o atributo que se mide. Las variables de ingreso, edad, número de hijos, etc. son ejemplos de este tipo de escala. El nivel de medición de razón se aplica tanto a variables continuas como discretas.

## 1.3 - Los Métodos Cuantitativos y la Estadística. Usos de los en las Ciencias Sociales

A medida que la Ciencia progresa, sus teorías se van haciendo más y más matemáticas en la forma. Hay una relación positiva entre el progreso de una Ciencia y el grado de su desarrollo matemático.

No es necesario que el investigador en Ciencias Sociales sea un especialista en áreas matemáticas concretas, lo verdaderamente importante es que sepa acercarse con mentalidad matemática a los problemas que se le plantean. La mentalidad matemática se define como la comprensión del proceso lógico subyacente al razonamiento matemático: entender la estructura formal del modelo matemático y las condiciones que lo hacen posible.

Tiene que haber un compromiso de manera que se simplifique la realidad concreta lo menos posible, pero, a la vez, lo suficiente para que el modelo creado a partir de la realidad sea fácilmente manejable desde el punto de vista instrumental matemático.

Es necesario una buena cierta matemática para conocer la potencia y debilidad de las técnicas estadísticas y consiguientemente para saber usarlas con eficacia y a la vez con prudencia.

Para estudiar Estadística Matemática se necesita cálculo avanzado y álgebra de matrices, sin embargo tal madurez no es indispensable para comprender las bases de la Estadística Aplicada.

El sentimiento de satisfacción y tranquilidad que resulta de dominar un lenguaje lógico y no ambiguo compensa la ocasional ansiedad que se desencadena al descubrir que se ha expresado un absurdo explícitamente y a todas luces.

Desde un punto de vista matemático muchas de nuestras afirmaciones están incompletas, mal encuadradas o son imprecisas.

Pero, por otro lado, muchas de estas ideas pueden ser entendidas intuitivamente, y es mejor una comprensión intuitiva que ninguna comprensión en absoluto.

Es preferible que los ejemplos que se usen en la enseñanza sean hipotéticos, porque es más importante tener un problema simple y plausible que el estudiante pueda comprender y que ilustre el método claramente que otro que simplemente asombre al estudiante con nuestra sabiduría.

Las principales aplicaciones estadísticas en cualquier campo, no sólo el de las Ciencias Sociales, descansan sobre el hecho de poder hacer observaciones o experimentos repetidos, esencialmente, bajo las mismas condiciones. En algunas áreas de la investigación, los objetos o fenómenos observados bajo las mismas condiciones variarán sólo en pequeña medida (en las ciencias físicas, donde las observaciones controladas dan prácticamente los mismos resultados).

Pero, por otro lado, especialmente en las Ciencias Sociales, aunque el experimentador haga un esfuerzo sobrehumano para observar repetidamente bajo las mismas condiciones, se encontrarán diferencias entre las observaciones y las diferencias, ordinariamente, no serán despreciables.

La Estadística Matemática es una teoría acerca de la incertidumbre, la tendencia de los resultados a variar cuando observaciones repetidas se hacen bajo condiciones idénticas.

La Estadística es el estudio de fenómenos donde, bajo un mismo conjunto de condiciones, las medidas obtenidas presentan variabilidad, y por tanto resultados impredecibles a priori; es decir, existe incertidumbre asociada al conocimiento del objeto de estudio. Aceptado que la Estadística trata sobre la incertidumbre, cabe preguntarse si la naturaleza está determinada o, en realidad, la incertidumbre es inherente a la misma, y por tanto está indeterminada. Y si está indeterminada entonces la Estadística tratará sobre la misma esencia de la realidad empírica.

Definamos entonces la Estadística como aquella manera de pensar de la cual se deriva una forma de representar los sistemas y razonar sobre ellos, sobre una naturaleza que se muestra indeterminada. La Estadística puede considerarse una Ciencia que guía la extracción de conocimiento, e implica una manera de conceptuar cualquier problema donde la incertidumbre es inherente a la comprensión del objeto de estudio y, por lo tanto, nuestro discernimiento sólo puede ser probabilístico y expresado mediante leyes estadísticas.

Aunque la organización de la información, las transformaciones y la depuración de los datos no sean características esenciales de la Estadística, eso no implica que no puedan ser incluidas en una definición de la disciplina.

El objetivo de la Estadística como Ciencia es mejorar el nivel de vida de la sociedad. Estadística deriva de la palabra Estado, y etimológicamente significa recoger información para tomar decisiones de cómo repartir comida o trabajo.

La Estadística moderna se ocupa de la recolección, análisis e interpretación de información, tanto cuantitativa como cualitativa. Y los métodos estadísticos son particularmente útiles cuando hay variabilidad en la medición.

## 1.3.1- Los Métodos Cuantitativos y su Utilidad en las Ciencias Sociales

Un estadístico trabajando en el campo de las Ciencias Sociales se ocupa de las siguientes cuestiones:

- ¿qué datos se necesita recoger?
- ¿cómo se pueden usar los recursos disponibles más eficientemente para recolectar los datos?
- ¿cómo especificar un modelo matemático que describa el proceso que ha generado los datos?
- depuración y transformación de los datos
- ¿cómo presentar los datos de manera que transmitan sus rasgos más esenciales de una manera clara?
- ¿qué conclusiones se pueden extraer de los datos y cuál es el grado de incertidumbre de estas conclusiones?
- ¿qué acciones se deben tomar en base a las conclusiones extraídas de los datos?

En la actualidad la Estadística es probablemente una de las disciplinas científicas más utilizada y estudiada en todos los campos del conocimiento humano. Por ejemplo: en la Administración de Empresas se utiliza para evaluar la aceptación de un producto antes de comercializarlo, en Economía para medir la evolución de los precios mediante números índice o para estudiar los hábitos de consumo mediante encuestas, en Ciencias Políticas para conocer las preferencias de los electores antes de la votación mediante sondeos y así orientar las estrategias de los candidatos, en Sociología para estudiar las opiniones de los colectivos sociales sobre temas de actualidad, en Psicología para elaborar las escalas de los tests y cuantificar aspectos del comportamiento humano, en general en las Ciencias Sociales para medir la relación entre variables y hacer predicciones sobre ellas.

En las Ciencias Sociales la Estadística se estudia en tres secciones: la Estadística Descriptiva, la Estadística Inferencial y el Diseño Experimental. La Estadística Descriptiva sirve de herramienta para describir, resumir o reducir las propiedades de un conglomerado de datos al objeto de que se pueda manejar. La Estadística Inferencial se utiliza para estimar las propiedades de una población a partir del conocimiento de las propiedades de una muestra de ella. Y en tercer lugar, el diseño y análisis de experimentos se desarrolla para determinar y confirmar relaciones causales entre variables.

En la investigación la Estadística es importante porque:

- permite el tipo más exacto de descripción,
- fuerza a ser exactos y definidos en nuestros procedimientos y pensamiento,

- permite resumir nuestros resultados de una forma conveniente,
- permite extraer conclusiones generales,
- permite predecir, y
- permite analizar algunos de los factores causales que subyacen a eventos complejos.

Dentro del campo de la Psicología hay tres vertientes metodológicas (lo cualitativo, lo no experimental y lo longitudinal) que son el autentico punto de partida de las actuales líneas de desarrollo. Las ecuaciones estructurales permiten la modelación de la causalidad. La regresión logística, los modelos log-linear y el análisis de correspondencias se utilizan para el análisis de datos cualitativos. Las series temporales investigan el aspecto longitudinal.

La mayoría de las técnicas utilizadas en los cuasi-experimentos se deriva del modelo de la regresión múltiple, de modo que las hipótesis rivales son probadas una a una. Por el contrario, en los estudios aleatorizados, se estima exactamente un efecto y se eliminan otros, dado que existe garantía de que influyen por igual en el grupo experimental y en el grupo control.

La Estadística capacita para:

**Interpretar las puntuaciones individuales de los sujetos en el contexto de los grupos de los que forman parte:** a todos nos resulta fácil interpretar la talla o el peso de una persona como elevada, media o baja. Así, una puntuación individual, por ejemplo 80 Kg y 168 cm, nos son relativamente fáciles de interpretar en nuestro contexto de referencia. De una manera más o menos precisa, nos hacemos idea de cómo se encuentra esa persona en relación con los miembros del grupo del que forma parte.

Pero esta interpretación puede ser más precisa si conocemos determinadas características del grupo, tales como la “normalidad” de talla y peso en el grupo de edad o sexo del que forma parte. A esto nos ayuda la Estadística, indicándonos cuál es la media aritmética* del grupo y en cuánto se aparta de esa puntuación* el sujeto de que se trate (dispersión*).

Sin embargo, esto mismo no es tan fácil si hablamos de variables diferentes, como la estabilidad emocional, la autoestima, la inteligencia, la sociabilidad, el nivel de conocimientos. La posibilidad de “medir” estas variables, con las limitaciones a que haremos referencia más adelante, nos pone en una situación próxima, aunque no tan exacta como la anterior.

Por otra parte, medidas específicas, como los cuartiles, la edad mental, el cociente intelectual o las puntuaciones típicas nos permitirán una interpretación más técnica y precisa.

**Caracterizar los grupos con los que trabaje: una clase, un curso, los miembros de una profesión (médicos, abogados, fontaneros, jornaleros...), los sujetos que han**

**realizado una encuesta, etc.:** Como acabamos de decir, la interpretación de las puntuaciones individuales suele hacerse en el contexto de los grupos de los que esa persona forma parte. Las ideas políticas de un individuo en relación con la clase social a la que pertenece, la inteligencia en el contexto de su edad y sexo, las calificaciones de Inglés en un curso y carrera determinada, los niveles de colesterol, la talla o el peso, teniendo en cuenta la edad, el sexo o el grupo de riesgo de que forma parte... son ejemplos de datos que debemos ser capaces de interpretar y valorar.

Pues bien: la Estadística, mediante las medidas de posición*, dispersión* y forma*, nos informa de esas características grupales.

Con estos valores, quienes deban tomar decisiones o hacer interpretaciones especializadas (profesores, orientadores, psicólogos, sociólogos, economistas...) pueden hacerlo con mayor seguridad de acertar que desconociendo tales datos.

**Extraer información de tales características para la toma de decisiones de carácter profesional:** Los científicos, los estudiosos, los profesionales y, en general, las personas interesadas en los diferentes campos del saber, no acuden a la Estadística por sí misma sino por la utilidad que les proporciona la información que les ofrece.

Un profesional de la Educación encontrará información relevante para organizar las actividades en su aula, para atender a la diversidad de sus alumnos, para mejorar sus programas, para predecir (y tomar decisiones preventivas) sobre los alumnos con riesgo...

Un psicólogo podrá caracterizar a sus pacientes, diagnosticar síndromes, recomendar tratamientos...

Un sociólogo será capaz de orientar a los políticos, interpretar estados sociales, adelantarse a las crisis...

Un economista ayudará a la empresa a prevenir problemas, a identificar riesgos, a diseñar campañas atendiendo a los perfiles de los clientes...

Un médico podrá estar al tanto de la incidencia de ciertas enfermedades, de los riesgos de determinados medicamentos, de las peculiaridades de ciertos pacientes en relación con algunos fármacos...

En fin: la utilidad de la Estadística tiene que ver con su ayuda a los profesionales para tomar decisiones que les son propias.

Y todo lo anterior –dimensión práctica- no reduce sus aportaciones al puro avance del saber, interés primordial del científico en sus diferentes ámbitos del conocimiento, sino que lo engrandece.

Identificar relaciones existentes entre las puntuaciones obtenidas por los miembros de esos grupos en dos o más variables: El ser humano, como persona aislada o formando parte de grupos, se comporta de modos muy diversos como consecuencia de la interacción entre sus características y las del contexto en que vive y de las relaciones existentes entre unas y otros.

La identificación de las relaciones entre variables de la persona o de estas con características individuales o grupales aporta gran información al sociólogo, al economista, al psicólogo o al pedagogo, incluso al médico. En fin: es una rica información para los profesionales que trabajan con personas.

Pues bien: la Estadística nos permite conocer si ciertas variables están relacionadas con otras o no, esto es: si varían conjuntamente (co-varían) o son unas independientes de otras.

Por ejemplo, si la inteligencia está o no relacionada con la clase social, si la autoestima se relaciona con la introversión, si agresividad y seguridad en sí mismo son independientes o están relacionadas...En estadística, representamos por lo general la correlación* con el símbolo r xy , esto es: la correlación entre las variables X e Y (por ejemplo: entre inteligencia y rendimiento académico, entre pobreza y analfabetismo...)

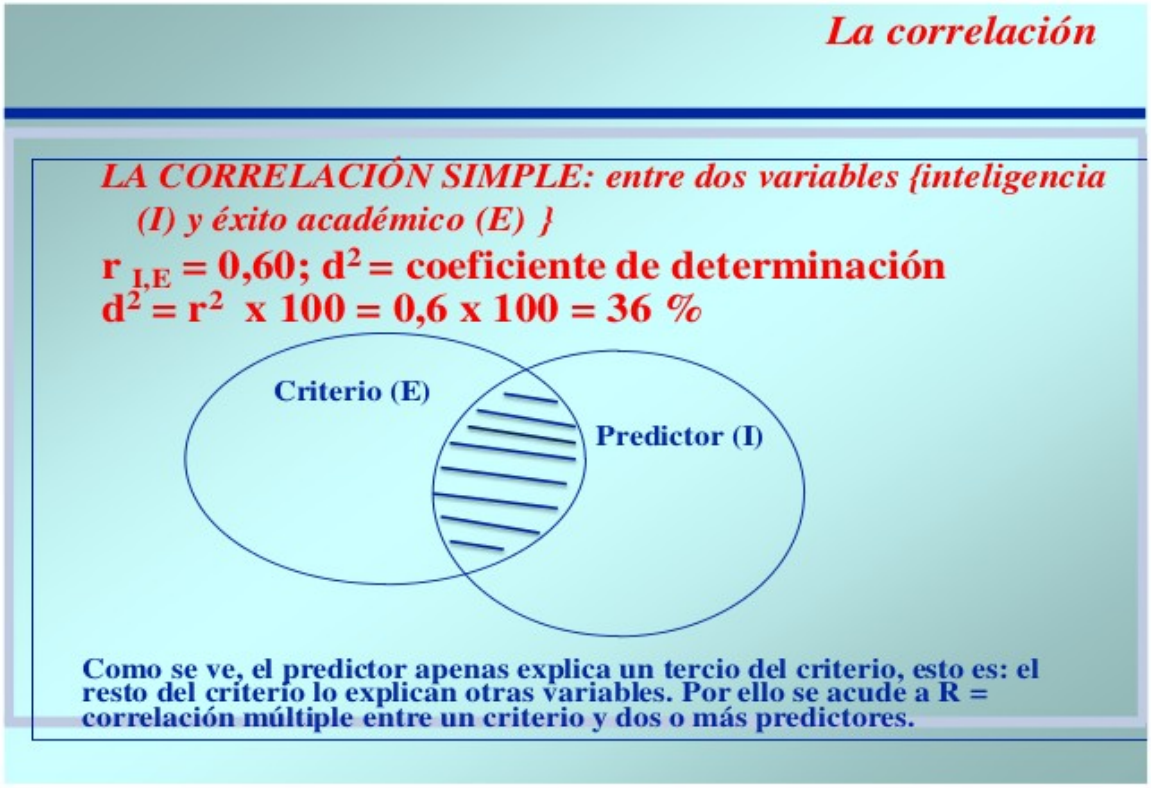
Esta información es fundamental para identificar las variables sobre las que poder intervenir cuando se desea modificar –positiva o negativamente- otra variable. Por ejemplo, conocer las variables que están ligadas (relacionadas) con la autoestima, nos ayuda a intervenir sobre aquellas para modificar esta. Sabiendo cómo se relaciona la motivación con el rendimiento, podemos incidir sobre aquella para elevar este; conocer la relación entre el dinero en circulación y el grado de inflación ayuda al político a tomar las medidas pertinentes, etc.

La Estadística nos informa sobre estas variables, sobre el tipo de relación (positiva: elevando los valores de una se elevan los de la otra, y viceversa) o negativa (elevando los valores de una disminuyen los de la otra y al contrario) y sobre su intensidad (perfecta, imperfecta –lo habitual, más o menos elevada- o nula).

Es más: a través de ciertas propiedades de las correlaciones podemos predecir, bien es verdad que con márgenes de error y determinados niveles de probabilidad*, lo que ocurrirá en una variable conociendo los valores obtenidos en otra. Por ejemplo: un orientador puede predecir, asumiendo cierto riesgo de equivocarse, qué alumnos suspenderán al final de curso en Estadística, a partir de una variable relacionada con ella, como son ciertas competencias matemáticas. Lógicamente, la intervención irá destinada a evitar que se cumpla la predicción. El margen de error ocurre en toda predicción, como la del tiempo, de la evolución de una enfermedad, de la famosa “prima de riesgo”, de las actitudes hacia los extranjeros, etc.

En la figura 1 podemos apreciar de forma intuitiva cómo un predictor como la inteligencia mantiene una relación con el éxito académico de aproximadamente 0,60, lo que viene a representar que explica poco más de una tercera parte del criterio (zona rayada de la figura 1.a).

Figura 1 Representación intuitiva de la correlación simple entre dos variables




## 1.3.2- Aplicación de las Propiedades de los Modelos Estadísticos

Creo que todo lo anterior ya justifica el estudio y dominio de los contenidos de la asignatura. Sin embargo, y aunque lo que viene a continuación exige ya una cierta base, las principales utilidades de la Estadística, están ligadas a su modalidad inferencial a la que solo haremos una breves referencias.

En esencia, esta parte de la Estadística trata de ir más allá de los datos empíricos, datos obtenidos mediante instrumentos como los test, los exámenes, los cuestionarios, las encuestas, la observación, las entrevistas... Como hemos anunciado, con ellos hemos podido interpretar una puntuación individual*, caracterizar un grupo o averiguar si se dan o no relaciones entre variables. Ahora la cuestión es más compleja: con los valores medidos a los integrantes de esos grupos, ¿podemos ir más allá y utilizarlos para interpretar las puntuaciones de otros sujetos que forman parte de grupos con las mismas características?

Veamos: Lo normal es que en Estadística trabajemos con los valores obtenidos por un grupo de sujetos de una edad, sexo, curso, carrera, raza, clase social, ideología, religión, ... en variables como actitudes, conocimientos, técnicas de estudio, autoconcepto, locus of control, nivel de pobreza... pero el interés del investigador es que, a partir de ellos, se puedan aplicar –con cierta prudencia y admitiendo márgenes de error- al conjunto de sujetos de la misma edad, sexo, curso, carrera, en la variable estudiada.

Los primeros valores, denominados estadísticos*, se "miden" en conjuntos denominados muestras*; los segundos son estimados para el conjunto total, denominado población*; los valores estimados se denominan parámetros*. Estimar es tanto como atribuirle un valor mediante procedimientos técnicos; no obstante, cualquier estimación está sujeta a errores que deben ser tomados en consideración, como lo hace la Estadística (error de estimación).

Pues bien: para poder hacer tal cosa, la Estadística aplica a los datos muestrales las propiedades de ciertos modelos*, para lo cual lo primero es decidir si a aquellos se les puede aplicar el modelo y sus propiedades.

Podemos entender esto fácilmente. No creo que nadie haya visto jamás en la realidad un cono perfecto; sin embargo, todos identificamos los volcanes –pensemos en el Teide- con una forma cónica. Admitiendo que la realidad nos presenta objetos cónicos –más o menos cercanos al cono ideal- podemos aplicar a tales objetos reales las propiedades del cono; del mismo modo podríamos actuar con el prisma, con la pirámide, con la esfera..., y calcular así la superficie y el volumen de un objeto piramidal, prismático o esférico.

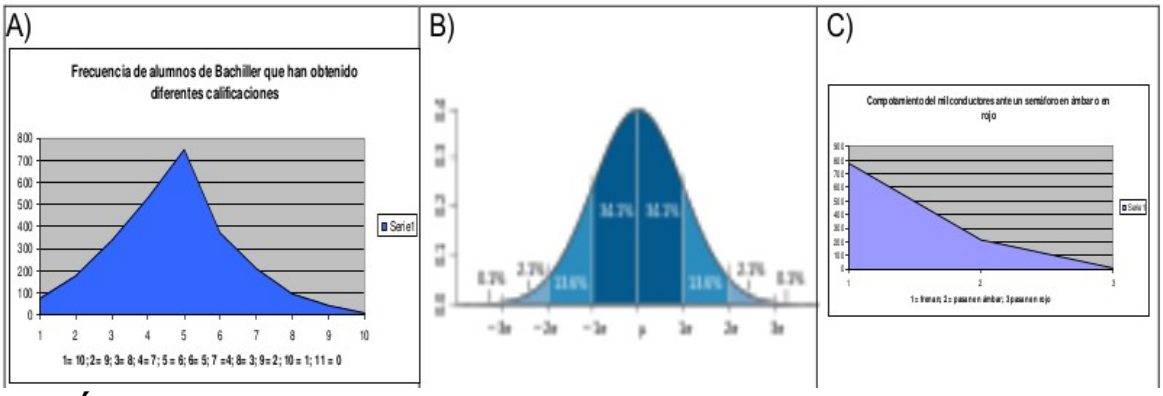
Un caso similar y sencillo en nuestro ámbito; todos conocen o han oído hablar de la denominada curva normal de probabilidades* o campana de Gauss. En sí misma es un modelo*, por tanto, algo ideal: no encontrarán en la naturaleza ninguna realidad igual a esa campana, pero sí datos más o menos próximos a ella. Pues bien: lo que se plantea es que si los datos reales se aproximan razonablemente bien al modelo –y esta es ya una cuestión estadística- sus propiedades puedan aplicarse a los datos reales, lo que supone un gran avance en el tratamiento de la información recogida.

En las figuras siguientes (Fig. 2) podrán apreciar el modelo normal –en el centro- y dos series de datos empíricos, más o menos cercanos al mismo. Decidir si se les pueden aplicar las propiedades de modelo normal es la cuestión que nos ayuda a resolver la Estadística:

Figura 2: Representaciones de tres series de datos, A, B y C

Tomando como normal la figura central, ideal, modelo teórico, parece claro que la primera figura se acerca más a ella que la tercera. Podemos descartar que esta sea “normal”, pero no tenemos seguridad de que la primera sí lo sea. La Estadística nos ayudará a decidirlo asignando a la decisión una determinada probabilidad* de estar en lo cierto. En caso afirmativo, podremos aplicarle las propiedades del modelo, al igual que utilizamos las del cono para estimar el volumen de una montaña cónica.

Obviamente, al aplicar tales propiedades somos conscientes de ciertas deformaciones; pero también es cierto que estas pueden deberse a que nuestra muestra* no era lo suficientemente representativa del conjunto de la población por problemas de tamaño y de forma de seleccionar sus componentes. En ese caso, esas deformaciones afectan a los datos empíricos pero estos serían más y más cercanos al modelo en la medida en que corrigiéramos tales deformaciones.




# CAPÍTULO DOS

# La Estadística Descriptiva

El siguiente capítulo busca describir los usos que los métodos cuantitativos y la estadística tienen en el desarrollo de las ciencias sociales.

Concluido el capítulo se espera que usted logre los siguientes objetivos:

- Comprender la lógica analítica propia de la estadística descriptiva e inferencial
- Conocer y comprender las distribuciones muéstrales más comunes en la investigación social y los elementos básicos de la inferencia estadística.
- Comprender principales estadígrafos descriptivos y el sentido técnico en el análisis de variables sociales.
- Elaborar e interpretar tablas y gráficos para la descripción de variables sociales.
- Organizar la información disponible familiarizándose con el lenguaje estadístico y el uso de matrices de datos.
- Adquirir la capacidad para interpretar resultados y relacionar los elementos de estadística con la investigación social.
- Obtener información clara y profunda que ge era nuevo conocimiento disciplinario.

## 2.1 - De Donde Provienen los Números en la Estadística

Imagina a un investigador en Psicología, Sociología, Pedagogía, Ciencias Económicas, Medicina en su oficina de trabajo. Ha recogido los protocolos de un cuestionario de opinión sobre sus respectivos objetos de estudio: actitudes hacia los inmigrantes, opiniones políticas, motivación, niveles de renta, incidencia de un virus... que se acumulan sobre su mesa de trabajo. Puede tener 100, 200, 1000 o más cuestionarios.

A continuación encontrará una serie de enunciados relacionados con la calidad de la educación, seguidos de un espacio destinado a recoger el grado de importancia que, a su juicio, reciben de hecho en el centro educativo en el que presta sus servicios. Su valoración se expresa entre 1, la mínima, y 4, la máxima. En la siguiente columna deberá marcar si, a su juicio, es manifiestamente necesario introducir mejoras en tal aspecto.

| ITEMS | Importancia que se le concede en su centro: | | | | Necesidad de mejora | |
|---|---|---|---|---|---|---|
| | 1 - | 2 | 3 | 4 + | SI | NO |
| ***I. UN PROYECTO EDUCATIVO DEL CENTRO:*** | | | | | | |
| *1. Que incorpore los valores morales y sociales de consenso* | | | | | | |
| *2. Que integre el carácter propio en el caso de los centros concertados y privados* | | | | | | |

¿Cómo pasamos de esos protocolos a los números con los que trabaja la Estadística*?

Esos cuestionarios pueden contener 10, 20, 50... preguntas o ítems que los sujetos pueden valorar, por ejemplo, asignando un 1 si responde SI, y un 0 cuando marca el NO.

Sumando los valores marcados llegamos a una puntuación con el número de “síes” y de “noes”. Del mismo modo, dado que es posible marcar entre 1 y 4 la valoración de cada ítem, según la importancia concedida a cada uno de los enunciados, podremos obtener la puntuación* total de cada persona consultada y hasta la tendencia del grupo, merced a una medida tan conocida como es la media aritmética.

Al final, cada uno de tales protocolos se ha convertido en un número, número con el que trabajará la Estadística. Este número se denomina puntuación directa* y suele representarse por X i , esto es: puntuación directa o bruta del sujeto i.

Recuerde: la puntuación directa de un sujeto cualquiera -también llamada “bruta”- en un instrumento de recogida de datos, se representa por X i y se lee: puntuación directa del sujeto i.

Los números que manejamos nacen de pesar, medir o contar los “objetos” más diversos. Objetos como la talla, el peso, la edad, la inteligencia, la asertividad, la renta “per cápita”, la autoestima, el rendimiento académico, la masa muscular, el “ranking” de un país en los Juegos Olímpicos (o en las pruebas PISA)... son objetos de medida y, mediante los instrumentos adecuados, dan lugar a números.

### 2.1.1 - Diferentes Tipos de Números

Parece claro, no obstante, que, por encima de la apariencia externa de los números todos son iguales en su apariencia- en realidad son muy diferentes unos de otros. Los 80 cm. de talla de un niño poco tienen que ver con los 80 puntos en una prueba de inglés, o los 80 Kg de peso, o el puesto 80 al llegar a meta, o el número de sujetos -80- que son admitidos a un concurso, o los 80 centígrados alcanzados por un horno, o los 80 puntos obtenidos en una prueba de autoestima, o...

Cuando hablamos de talla, o de peso, estamos ante los números plenos, además de fiables y válidos si han sido medidos con cuidado y utilizando un metro y una balanza fiables. Y ello se debe a que contamos con unidades de la misma naturaleza que el objeto a medir: el centímetro para la talla o el kg. para el peso, y damos por hecho que quien mide lo hace con seriedad.

Aparentemente, los 80o centígrados del horno son de la misma naturaleza que los anteriores, pero no es así, por una sencilla razón: si antes el 0 significaba que no tenemos delante a nadie (porque nadie pesa 0 gr. ni mide 0 cm.), ahora, como todos sabemos, por debajo de 0º sigue habiendo temperatura: -3o, -15o, etc.

También se puede asignar ese número a un corredor de maratón que ha llegado a meta en el puesto 80. Pero aquí no medimos la distancia recorrida sino el orden de llegada; y puede haber diferencias notables en minutos o segundos entre las llegadas de los diferentes atletas. Puede ocurrir que entre el primero y el segundo apenas haya un par de segundos pero que entre este y el tercero haya más de un minuto, y que, más adelante, entre un grupo de corredores prácticamente en el mismo tiempo pero distanciados del anterior en 15 o 20 minutos. Pero eso no importa si lo que medimos es el orden en que entran.

Y, obviamente, el número de 80 de admitidos a un determinado concurso, 35 varones y 45 mujeres por ejemplo, solo nos indica las veces que personas varones y personas mujeres han sido seleccionadas, sin más.

Un mismo valor numérico puede representar objetos reales o empíricos muy diferentes; según sean estos objetos, al número que los representa se les podrán aplicar unas u otras propiedades de los números y sus correspondientes operaciones matemáticas.

### 2.1.2 - Escalas de Medida

Pues bien: cada uno de esos 80 representa un tipo diferente de número, propios de diferentes niveles o escalas de medida: los de razón o cociente, en el primer caso (talla, peso), permiten todo tipo de operaciones aritméticas; los de intervalo, en el segundo (grados centígrados), con los que podemos establecer ciertas operaciones pero no otras (no conviene entrar aquí en detalles); los de orden (puesto ocupado al llegar a meta) nos indican solo lo que es mayor o menor, anterior o posterior, más o menos intenso..., pero no podemos operar con ellos de otra manera; o de tipo nominal, que solo nos indican que algo es igual o diferente que otro algo, pero no podemos hacer operaciones con ellos: asignar un 1 a los varones y un 2 a las mujeres no significa que estas sean más –o aquellos, menos- que los varones, sino, simplemente, diferentes. No tendría sentido, por lo tanto, sumar el número de unos y el doses e intentar calcular la media.

Ahora bien: observe el lector algo importante: hay objetos fácilmente medibles, porque están abiertos a nuestros sentidos (talla, peso, edad...) y porque tenemos unidades de medida de la misma naturaleza: cm, gr., año...

Pero hay otros que son, en realidad, objetos cuya misma naturaleza no conocemos y, por ello, tenemos que definirlos previamente. Pensemos en la inteligencia, en la asertividad, en la autoestima, en la opinión... y hasta en el rendimiento académico. Nadie ha visto la inteligencia, pero sí a personas inteligentes, o asertivas, o con baja autoestima, o con rendimiento adecuado...

Para “medirlas” debemos, en primer lugar, definirlas. A Binet se debe una famosa frase cuando se le preguntó ¿qué es la inteligencia? Su respuesta fue tan contundente y clara como discutible: Inteligencia –dijo- es lo que mide mi test. Podríamos decir que, a partir de otros autores de tests la inteligencia, llegaríamos a diferentes medidas de este rasgo humano (Y así es, por cierto). Y lo mismo pasará con el rendimiento académico: diferentes exámenes darán lugar a diferentes resultados, diferentes medidas. Y nada digamos si hablamos de asertividad, de autoestima, de esquizofrenia, etc. Una vez definido el objeto, debemos encontrar manifestaciones acordes con

la definición o elaborar reactivos que se consideren evidencias del mismo. Estos reactivos o estas manifestaciones se convierten en ítems del instrumento de medida. Por lo general, a esta traducción de la definición a ítems se la llama definición operativa u operacional.

Aquí nos encontramos con serios problemas para encontrar una regla de medida y su correspondiente unidad de medida y, por tanto, para asignar valores numéricos a la realidad medida. He aquí un problema que tendrán que conocer en su estudio de la asignatura.

Medir determinados objetos de los ámbitos en que trabajamos – Educación, Economía, Medicina, Psicología, Sociología...- implica definir el objeto a medir, encontrar manifestaciones de tal objeto o reactivos adecuados y decidir la regla de medida, la regla que nos permitirá atribuir un valor a cada manifestación o reactivo unidad de medida).

Nosotros dejamos constancia de tal problema, señalando las limitaciones que ello representa para los números que utilizamos, en particular:

a) Para las operaciones matemáticas que están justificadas con tales números

b) Para su fiabilidad: los números obtenidos en una ocasión pueden variar en otra

c) Para su validez: podemos estar midiendo una cosa que no es por completo la cosa deseada.

## 2.1.3 - Variables y Escalas de Medida

Ciertos “objetos” no presentan manifestaciones diferentes. Se les denomina constantes. Sin embargo, otros si las tienen, tales y se les denomina variables*; tal es el caso del sexo, masculino o femenino; del estado civil: soltero, casado, divorciado o viudo; de los grados universitarios: Pedagogía, Psicología, Sociología...; junto a estos, en otros casos los objetos a medir admiten valores que difieren en cantidad. A las primeras, las denominamos variables cualitativas, y se miden con números propios de escalas nominales mientras las segundas se conocen como cuantitativas*, y admiten números ordinales, de intervalo o de razón.

Las cualitativas pueden presentar dos categorías –dicotómicas*, como en el caso del sexo- o más, en cuyo caso hablamos de cualitativas politómicas*, como ocurre con el estado civil.

Algunos autores hablan de variables cuasi-cuantitativas*, en las que la cantidad solo puede apreciarse en términos de orden, por lo que son propias de escalas ordinales. Una variable de este tipo es la escala de dureza de los cuerpos en la que cada cuerpo está por delante o detrás de otro según que le raye o sea rayado por él (A Friedrich Mohs se debe una escala de 10 niveles de dureza que van del talco, el más blando, al diamante, al que solo puede rayar otro diamante).

A su vez, estas variables cuantitativas se dividen en discretas* (variables continuas que solo admiten valores enteros, como número de hijos o de alumnos) y continuas*, como es la edad, el peso, la talla...donde podemos asignar todos los valores intermedios si disponemos de los instrumentos adecuados (una balanza de precisión, por ejemplo; o un cronómetro, como el utilizado en las pruebas olímpicas de atletismo). En el cuadro 2 aparece una clasificación de las variables.

Figura 3: Clasificación de las Variables[1]

| TIPOS DE VARIABLES | ESCALA DE MEDIDA |
|---|---|
| Cualitativas | Nominales<br>• Dicotómicas: sexo<br>• Politómicas: estado civil, clase social, grado universitario... |
| Cuasi-cuantitativas | Ordinales:<br>• Escala de dureza<br>• Rangos o puestos<br>• Clasificación de los terremotos |
| Cuantitativas | Continuas: edad, talla, peso<br>Discretas: número de alumnos |

En algunos de estos casos caben transformaciones; así, una variable cuantitativa continua puede ser “tratada” como discreta, prescindiendo de algunos de los valores

1 Fuente: Adaptación Propia de “Estadística Aplicada a las Ciencias Sociales. Ramón Pérez Juste. Universidad Nacional de Educación a Distancia

posibles; por ejemplo: podemos tomar valores de edad de los alumnos quedándonos con los años y obviando los meses, o tomando años y meses, obviando semanas, días...

Una dificultad añadida se da en el caso de datos cualitativos, como los surgidos de entrevistas, que se desea tratar.

### 2.1.4 - ¿Que nos dicen los números?

Cuando un sociólogo hace una encuesta sobre intención de voto, obtiene determinados valores que suele traducir a porcentajes para su interpretación.

Cuando un psicólogo aplica un test de autoestima a un grupo de alumnos, asigna a cada uno una puntuación; esta puntuación oscila entre un piso y un techo (mínima y máxima), cuyos valores dependen de la regla de medida y de su correspondiente unidad de medida (por ejemplo: un punto por cada respuesta positiva).

Cuando un profesor propone a sus alumnos un examen, asigna a cada uno de ellos una calificación que, del mismo modo, depende del número de ítems o preguntas y de la regla de medida: por ejemplo: número de respuestas acertadas menos número de errores, partido por el número de alternativas ofrecidas menos 1, fórmula habitual para la calificación de las pruebas objetivas (ecuación 1):

Puntuación ($X_i$)

$$Ecuación\ 1:\ \ Xi = \sum_{i=1}^{N} A - \frac{E}{a-1}$$

Obviamente, “medir” el rendimiento con un examen de desarrollo multiplica los problemas para decidir cuál es la unidad y, en consecuencia, cuál es el valor a asignar a cada examen.

Veamos algunos casos:

Un cuestionario de 40 preguntas (ítems) en que el encuestado puede marcar SI, NO, NO SÉ.

Una prueba objetiva en la que el profesor decide valorar solo las preguntas bien resueltas.

Otra prueba objetiva en la que el profesor aplica la fórmula anterior, restando los errores teniendo en cuenta que las alternativas ofrecidas son 4.

Una escala de actitud en que para cada ítem el consultado debe marcar su posición entre 1 (mínimo) y 7 (máximo).

Parece evidente que si en cada una de esas situaciones una persona obtiene 25 puntos, tal valor no puede interpretarse del mismo modo ni puede significar lo mismo.

Debemos tomar conciencia de la importancia que cobra la regla de medida; con ella atribuimos valor –números- a la información recogida con los diferentes instrumentos.

Ahora bien: conviene reflexionar sobre el carácter arbitrario que, en muchas ocasiones, tiene la decisión sobre tal regla, y lo que ello representa para el trabajo con tales números.

## 2.1.5 - ¿Cómo Interpretar los Resultados?

**Puntuaciones Individuales**

Si nos interesa una puntuación individual (representada por X i: puntuación directa del sujeto i) solo nos hacemos una idea de dos maneras: conociendo los valores extremos o poniendo su puntuación en relación con el grupo. Por ejemplo, 9 puntos es un sobresaliente si la prueba se puntúa sobre 10, pero es una puntuación muy baja si lo es sobre 50. Y también es una puntuación baja si la mayoría de las puntuaciones obtenidas por los sujetos del grupo están por encima de los 30 puntos.

Otra forma de interpretar las puntuaciones es a través de determinadas transformaciones de las puntuaciones individuales directas, a ciertas medidas, como puede ser un cuantil.

Entre los cuantiles, los más usados son el cuartil, el decil y el centil o percentil; estas medidas nos indican la posición de un sujeto cuando el grupo se ordena en cuatro, diez o cien partes. Así, estar en el cuartil 1 (Q 1 ) es encontrase entre el 25 % inferior del grupo; hallarse en el decil 7 (D 7 ) equivale a superar al 70 % del grupo, y obtener una puntuación equivalente al centil o percentil 78 (C 78) viene a ser superar al 78 % del grupo.

Medidas individuales son, también, la puntuación de desviación –puntuación diferencial- representada por x, que no es sino su separación –negativa o positiva- en relación con la media del grupo (x i = X i – Media).

Así, un sujeto cuya x i = -2 nos indica, de entrada, que puntúa por debajo de la media aritmética (signo negativo) y, en concreto, que se aparta dos puntos de la misma.

La Edad Mental (EM) es, también una puntuación individual, y su cociente con la edad cronológica (EC) otra diferente, el Cociente Intelectual (CI): CI = EM / EC. La EM indica que una persona tiene una inteligencia propia de una determina edad. Por ejemplo, si un niño tiene EM = 9, estamos diciendo que su desarrollo intelectual equivale al de un niño ideal de 9 años; claro está: si tal niño tiene en realidad 12, estamos afirmando que tiene retraso mental, pero si tuviera 8, la interpretación es que es un niño con desarrollo intelectual por encima del propio de su edad.

Para una mejor interpretación se ha desarrollado el CI, por lo general multiplicado por 100. De esta forma, un niño de 6 años y EM de 6, tiene un CI = 1, o de 100, si lo multiplicamos por 100.

Tanto ese 1 como el 100 nos informan de un niño cuyo desarrollo intelectual es normal, apropiado a su edad cronológica.

Otra medida, que exige previamente el cálculo de medidas de grupo (a las que nos referiremos en seguida), es la puntuación z i ; esta puntuación individual es el cociente entre la puntuación diferencial (x i ) de cada persona (puntuación directa menos la media del grupo) y la desviación típica de este. En resumen, la z i indica en cuántas desviaciones típicas del grupo se aparta un sujeto cualquiera de la media del mismo

(ecuación 2). Para entendernos: lo mismo que hablando de distancias utilizamos como unidad el Km., en este caso tomamos con unidad el valor de la desviación típica.

Entenderemos mejor esto al hablar en su momento de la curva normal de probabilidades*. En efecto: con ella presente sabremos que z i = 0 representa a un sujeto normal, en la media; que z i = -1 es propia de un sujeto que supera al 34 %, mientras que z i = + 1 lo hace con el 84 % aproximadamente. Clarificaremos estos conceptos más adelante.

$$Ecuación\ 2: \quad z_i = \frac{x_i}{s} : \frac{(X_i - Media)}{s}$$

Medidas individuales son aquellas que se refieren a un solo sujeto; como se ha indicado, su puntuación directa* se representa por X i . Para interpretar este valor podemos acudir a x i o puntuación diferencial * con respecto a la media; a z i , que indica en cuántas desviaciones típicas se aparta el sujeto de la media aritmética* del grupo; o a los diversos cuantiles (Q, D o P).

**Puntuaciones Grupales**

Si nuestro interés es interpretar las puntuaciones del grupo, y este es pequeño, no resulta difícil hacernos una idea de cómo es ese grupo; sin embargo, cuando es grande, ver las puntuaciones tal y como van apareciendo al ser calificados los exámenes o valorado un test, o recogidas las puntuaciones de un cuestionario... se convierte en algo complejo: los números parecen una realidad confusa e informe, como se aprecia en el siguiente conjunto de datos:

**Serie de datos 1:**

72, 87, 95, 88, 79, 69, 55, 54, 69, 77, 88, 60, 64, 60, 88, 77, 67, 75, 75, 52, 52, 67, 77, 95, 87, 60, 95, 86, 77, 67, 85, 51, 51, 67, 77, 85, 94, 64, 64, 50, 94, 93, 85, 76, 64, 75, 91, 82, 85, 62, 62, 77, 82, 91, 90, 80, 85, 82, 110, 75, 62, 62, 75, 72, 80, 62, 94, 90, 67, 85, 54, 60, 90, 72, 80, 22, 79, 89, 57, 89, 79, 8, 57, 77, 71, 76, 89, 91, 54, 70, 94, 79, 57, 55, 70, 89, 70, 88, 26, 10

N = 100

Ante estos hechos, la Estadística* nos ayuda mediante la organización de los datos, en particular a través de su ordenación y reducción o simplificación.

**Organización de los Datos**

La primera operación que suele realizarse es la de ordenar los números, las puntuaciones. Una operación tan sencilla como esa nos permite conocer:

- Las puntuaciones extremas; puestas las puntuaciones individuales en relación con las extremas posibles del cuestionario, de la prueba objetiva,... ya nos ofrecen una información interesante.

- La continuidad o no de las mismas, apreciando si se dan o no valores vacíos, huecos.
- La acumulación o no y en qué parte –superior, central o inferior- de la distribución de valores ordenados.

Veamos varios casos en una sencilla escala entre 0 y 10:

Caso a): 8, 6, 6, 6, 5, 3, 3, 3, 2, 2.

- Aquí apreciamos que no aparecen las puntuaciones extremas (9 y 10, 0 y 1)
- Que hay valores vacíos: 7 y 4
- Que se da una doble acumulación de puntuaciones, una en la parte superior y otra en la inferior (6 y 3, repetidos en tres ocasiones)

Caso b): 9, 8, 7, 6, 5, 5, 5, 4, 3, 2

En esta ocasión tampoco el grupo presenta puntuaciones a lo largo de todo el recorrido (falta el 10, el 1 y el 0), pero no hay huecos (hay mayor continuidad que en el caso anterior), se da una acumulación en el centro (valor 5) y una notable simetría en torno a la puntuación central.

Caso c)

1, 1, 1, 1, 1, 1, 1, 1, 1, 1

5, 5, 5, 5, 5, 5, 5, 5, 5, 5

9, 9, 9, 9, 9, 9, 9, 9, 9, 9

En estos tres ejemplos la distribución es uniforme: todos los sujetos alcanzan la misma puntuación, pero en el primero todas son bajas y en el tercero todas elevadas, frente a al segundo, de puntuaciones medias.

Comparemos ahora estas dos:

5, 5, 5, 5, 5, 5, 5, 5, 5, 5

10,10, 10, 10, 10, 0, 0, 0, 0, 0

Como vemos, estamos ante la máxima homogeneidad y la máxima heterogeneidad respectivamente. Cuando calculemos las medias aritméticas, veremos que en ambos grupos la media es la misma (5), pero un profesor que tuviera que trabajar con uno u otro grupo debería actuar claramente de formas bien distintas.

Los casos anteriores nos ilustran sobre el valor de una operación tan simple como es la ordenación –creciente o decreciente- de las puntuaciones. Fácilmente se comprenderá que esa utilidad es mucho mayor si, en lugar de los 10 casos, tuviéramos ante nosotros 100, 400, 1000...

Pero en ciertos casos, cuando el tamaño es elevado –pongamos 100 o más casos- la ordenación es laboriosa y su utilidad queda limitada, como fácilmente se desprende de los datos anteriores que utilizaremos más adelante.

La forma más sencilla de hacernos cargo de ciertas características de un grupo consiste en ordenarlos de forma creciente o decreciente. Esta sencilla acción nos permite apreciar su recorrido (diferencia entre las puntuaciones extremas), si se da o no continuidad a lo largo del mismo, su dispersión o variabilidad o la forma y el lugar en que se agrupan las puntuaciones.

### 2.1.6 - La Reducción. Distribución de Frecuencias

Cuando el conjunto de casos es elevado, como en la anterior serie 1 (el valor de N es de 100) una forma de facilitar la interpretación es mediante la reducción del conjunto a otro menor. El caso más sencillo se da cuando se evita la repetición de las puntuaciones.

Estamos hablando de una distribución de frecuencias en la que, por un lado, tenemos las puntuaciones obtenidas (X i) y, por otra, las veces que cada puntuación aparece en el conjunto de casos (f i ).

Así, con los casos a y b del apartado anterior podríamos reducirlos, quedando del siguiente modo (Cuadro 4):

Caso a): 8, 6, 6, 6, 5, 3, 3, 3, 2, 2.

Caso b): 9, 8, 7, 6, 5, 5, 5, 4, 3, 2

Figura 4: Reducción de Datos Originales a una Distribución de Frecuencias[2]

| Caso a) | | Caso b) | |
|---|---|---|---|
| $X_i$ | $f_i$ | $X_i$ | $f_i$ |
| 8 | 1 | 9 | 1 |
| 6 | 3 | 8 | 1 |
| 5 | 1 | 7 | 1 |
| 3 | 3 | 6 | 1 |
| 2 | 2 | 5 | 3 |
| | | 4 | 1 |
| | | 3 | 1 |
| | | 2 | 1 |

Supongamos que hemos realizado un examen, consistente en una prueba objetiva, a un total de 30 alumnos. Las calificaciones, a fin de que sean fácilmente comprensibles, las hemos reducido a la escala 0 - 10, habitual en el ámbito académico. Cabe pensar que estas calificaciones se puedan considerar ordinales e, incluso, de intervalo, dado que disponemos de una unidad de medida razonablemente precisa. He aquí los datos (Serie 2): X i : 9, 7, 7, 4, 5, 6, 7, 3, 1, 8, 8, 9, 3, 4, 10, 6, 3, 4, 8, 7, 1, 3, 2, 5, 7, 5, 4, 5, 8, 2

2Fuente: Adaptación Propia de “Estadística Aplicada a las Ciencias Sociales. Ramón Pérez Juste. Universidad Nacional de Educación a Distancia

Podemos hacer la ordenación de mayor a menor, que ya nos informará de las características de este grupo de alumnos: X i : 10, 9, 9, 8, 8, 8, 8, 7, 7,7, 7, 7, 6, 6, 5, 5, 5, 5, 4, 4, 4, 4, 3, 3, 3, 3, 2, 2, 1, 1

Como vemos:

Únicamente nos falta la puntuación 0

Se da continuidad de las puntuaciones

Apreciamos una mayor concentración de puntuaciones elevadas

Si reducimos el conjunto de datos, podemos apreciarlo más claramente. Para ello basta construir una distribución de frecuencias. Nótese que entre la serie anterior y la siguiente no se dan sino diferencias de forma pero no de contenido:

Figura 5: Distribución de frecuencias correspondiente a la serie 2[3]

| $X_i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | N |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $f_i$ | 2 | 2 | 4 | 4 | 4 | 2 | 5 | 4 | 2 | 1 | 30 |

Como vemos, las 30 puntuaciones han quedado reducidas a 10 diferentes y la acumulación de las puntuaciones repetidas, tomadas como frecuencias, nos permite una mayor y más fácil comprensión de las características del grupo:

a) Heterogéneo

b) Continuo: no se aprecia discontinuidad entre las puntuaciones

c) Con tendencia hacia las puntuaciones más elevadas: si tomamos el 5 como suficiente o aprobado, 18 de las 30 lo alcanzan y lo superan.

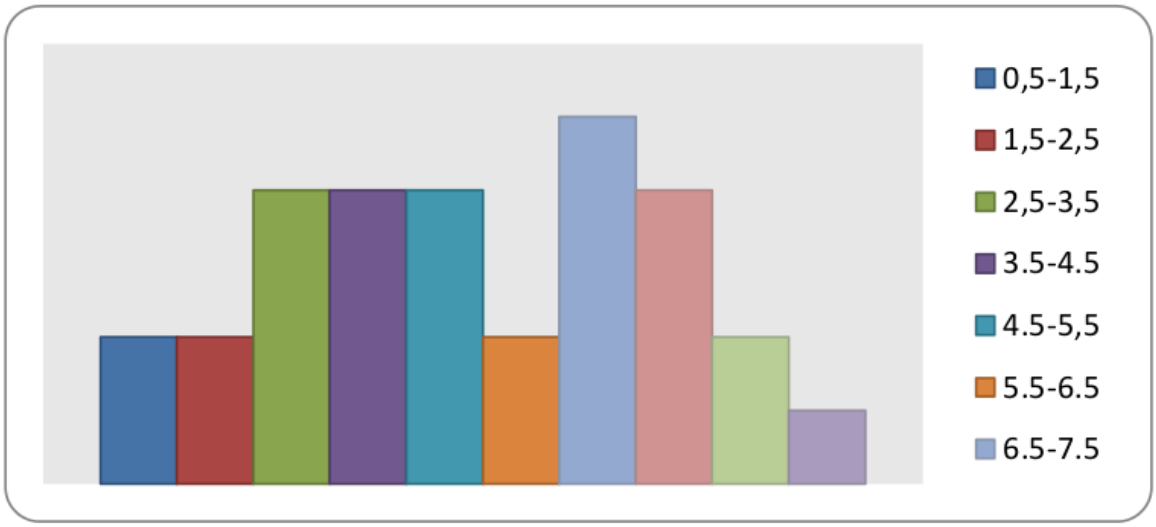
d) Además, en lugar de concentrarse el mayor número de casos en lo que podríamos llamar lo “normal”, esto es: en torno al 5, comprobamos que, además de los de puntuación muy alta (el 9 y el 10), lo que predomina son las puntuaciones elevadas (los notables).

Estas características son de gran relevancia para un profesor que deba atender las diferencias entre sus alumnos, o para un orientador que tenga que trabajar la autoestima de los mismos.

Los datos anteriores pueden presentarse de forma intuitiva mediante una representación gráfica conocida como histograma* (figura 3) consistente en un eje de coordenadas, con los diferentes valores en el eje de abscisas y con las frecuencias en el de ordenadas.

3 Fuente: Adaptación Propia de “Estadística Aplicada a las Ciencias Sociales. Ramón Pérez Juste. Universidad Nacional de Educación a Distancia

Figura 6: Histograma Correspondiente a los Datos de la Tabla 1




Una distribución original de datos, ordenada o no, puede reducirse por medio de una distribución de frecuencias; en ella se presenta una columna -o una fila- con las diversas puntuaciones, representadas por X i , y otra con las frecuencias -f i - o veces que cada puntuación se repite. La representación gráfica adecuada es el histograma.

Sin embargo, todavía es posible una reducción mayor de los datos, algo necesario cuando el rango o recorrido de las puntuaciones (diferencia entre los valores extremos, representado por R) es mayor, como ocurre con los siguientes datos, ya aludidos previamente (serie 1), en donde estamos ante 100 sujetos (N = 100) con puntuaciones que pueden ir de 0 a 130, presentados primero de forma “natural”, desordenada, y luego ordenados en forma decreciente:

Serie desordenada, con las puntuaciones X i según aparecen al investigador (Serie 1):

72, 87, 95, 88, 79, 69, 55, 54, 69, 77, 88, 60, 64, 60, 88, 77, 67, 75, 75, 52, 52, 67, 77, 95, 87, 60, 95, 86, 77, 67, 85, 51, 51, 67, 77, 85, 94, 64, 64, 50, 94, 93, 85, 76, 64, 75, 91, 82, 85, 62, 62, 77, 82, 91, 90, 80, 85, 82, 110, 75, 62, 62, 75, 72, 80, 62, 94, 90, 67, 85, 54, 60, 90, 72, 80, 22, 79, 89, 57, 89, 79, 8, 57, 77, 71, 76, 89, 91, 54, 70, 94, 79, 57, 55, 70, 89, 70, 88, 26, 10

N = 100

Serie ordenada en forma decreciente, correspondiente a los datos anteriores:

110, 95, 95, 95, 94, 94, 94, 94, 93, 91, 91, 91, 90, 90, 90, 89, 89, 89, 89, 88, 88, 88, 88, 87, 87, 86, 85, 85, 85, 85, 85, 85, 82, 82, 82, 80, 80, 80, 79, 79, 79, 79, 77, 77, 77, 77, 77, 77, 77, 76, 76, 75, 75, 75, 75, 75, 72, 72, 72, 71, 70, 70, 70, 69, 69, 67, 67, 67, 67, 67, 64, 64, 64, 64, 62, 62, 62, 62, 62, 60, 60, 60, 60, 57, 57, 57, 55, 55, 54, 54, 54, 52, 52, 51, 51, 50, 26, 22, 10, 8.

N = 100

El rango, en este caso es: R = 110 – 8 + 1 = 103 puntuaciones diferentes posibles. La mera ordenación ya nos permite ver el amplio recorrido de las mismas, con valores que, por una parte, se acercan a las puntuaciones más extremas (8, cerca del 0, y 110, próximo a la puntuación máxima de 130) .

Pero, por otra parte, si reducimos la serie a las puntuaciones directas (X i ) con sus correspondientes frecuencias (f i), como hemos hecho en el caso anterior, podemos apreciar que estamos ante una distribución todavía muy amplia todavía de no fácil apreciación de una forma global e intuitiva: nada menos que 35 valores:

Figura 7: Distribución de Frecuencias (Amplitud del intervalo = 1) Correspondiente a los Datos de la Serie 1

| | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $X_i$: | 110 | 95 | 94 | 93 | 91 | 90 | 89 | 88 | 87 | 86 | 85 | 82 | 80 | 79 | 77 | 76 | 75 | 72 | | | |
| $f_i$: | 1 | 3 | 4 | 1 | 3 | 3 | 4 | 4 | 2 | 1 | 6 | 3 | 3 | 4 | 7 | 2 | 5 | 3 | | | |
| $X_i$: | 71 | 70 | 69 | 67 | 64 | 62 | 60 | 57 | 55 | 54 | 52 | 51 | 50 | 26 | 22 | 10 | 8 | | | | |
| $f_i$: | 1 | 3 | 2 | 5 | 4 | 5 | 4 | 3 | 2 | 3 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | | | | |

Por ello es frecuente que la distribución tome la modalidad de intervalos, esto es: se trata de una distribución que nos indica cuantos casos (frecuencias: f i ) hay para un conjunto de puntuaciones que denominamos intervalos (I). Lógicamente, las frecuencias serán tanto más elevadas cuanto menor sea el número de intervalos. Por ello hay que decidir con prudencia cuántos intervalos

teniendo en cuenta el recorrido del conjunto y la amplitud que queremos dar a cada uno.

A tales efectos, debemos pensar que siempre que hagamos una distribución de intervalos vamos a “deformar” la distribución original en mayor o menor grado, ya que a todas las puntuaciones del intervalo las vamos a representar por una, la que ocupe el lugar central de cada intervalo (marca de clase, representada por X i, al igual que la puntuación directa). Sin embargo, cabe pensar que las deformaciones en un intervalo en un sentido tenderán a compensarse con las de otros intervalos en sentido contrario. Veamos.

En nuestro caso, la distribución oscila entre 8 y 110 puntuaciones; por tanto, hay (110 – 8) + 1 puntuaciones posibles; podemos hacer una distribución por intervalos; si tomamos la decisión de que su amplitud sea de 10 puntos, la distribución podría ser la siguiente (figura 8):

Figura 8: Distribución de frecuencias (amplitud del intervalo = 10) correspondiente a los datos de la serie 1.

| I | 1-10 | 11-20 | 21-30 | 31-40 | 41-50 | 51-60 | 61-70 | 71-80 | 81-90 | 91-100 | 101-110 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $X_i$ | 5.5 | 15.5 | 25.5 | 35.5 | 45.5 | 55.5 | 65.5 | 75.5 | 85.5 | 95.5 | 105.5 | |
| $f_i$ | 2 | 2 | 0 | 0 | 1 | 16 | 19 | 25 | 23 | 11 | 1 | $\sum$ = 100 |
| $X_i f_i$ | 11 | 31 | 0 | 0 | 45.5 | 888 | 1244,5 | 1887,5 | 1966,5 | 1050,5 | 105.5 | $\sum$ =7230 |

Como se puede apreciar, esta distribución es mucho más manejable y hasta intuitiva; a simple vista apreciamos su gran heterogeneidad

Su discontinuidad en la parte inferior, con dos grandes huecos de puntuaciones carentes de sujetos (a partir de la puntuación 20 hasta la 40, ambas inclusive). Cabe pensar que los cuatro sujetos inferiores de los dos primeros intervalos de la distribución podrían considerarse ajenos al grueso de grupo.

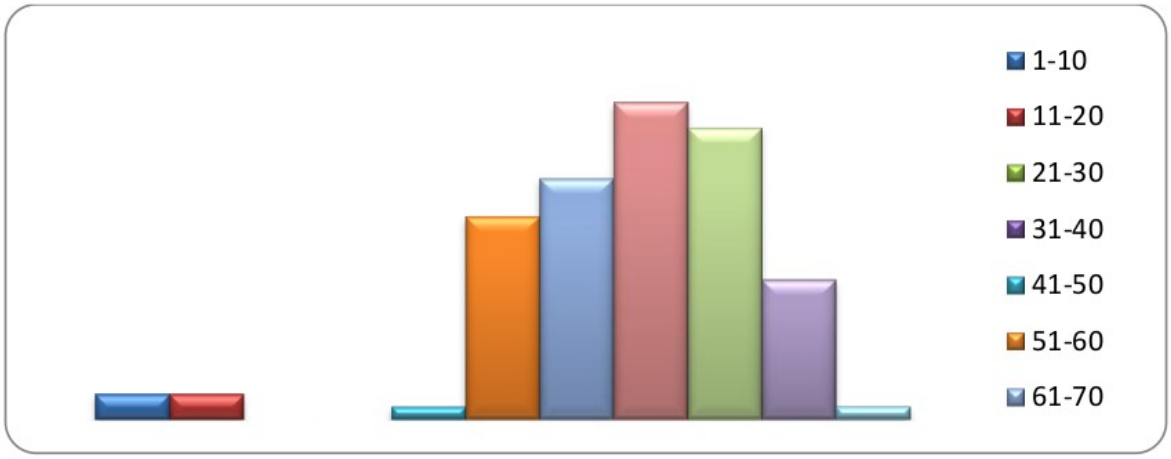
La tendencia a valores elevados, no solo por el caso que se encuentra en el intervalo superior sino porque las mayores frecuencias se sitúan claramente a la derecha de la misma

Debemos reconocer cierta distorsión. La más clara está en la puntuación superior, 110, que queda disminuida al ser representada por la marca de clase (105.5), al igual que la inferior, 8, que quedará representada por 5.5. Sin embargo, se acepta que en otros casos ocurrirá al contrario y que, en conjunto se compensan. En efecto, la puntuación 22 será representada por 25,5 y la 94 por 95.5.

Además, no debemos olvidar que, en general, estamos trabajando con números que no son totalmente fiables, que su fiabilidad no es total, por lo que la aparente pérdida de precisión no es tal si reconocemos esas limitaciones de los números que utilizamos.

Una representación gráfica de estos datos es el histograma, con una base proporcional a la amplitud del intervalo y una altura relaciona con su frecuencia (figura 9):

Figura 9: Histograma Correspondiente a los Datos de la Serie 1




Cuando el recorrido de la variable es muy amplio, es preferible acudir a una distribución por intervalos. En este caso, utilizamos una fila, o una columna para los intervalos y otra para las frecuencias. En teste caso, las frecuencias son las correspondientes a la amplitud de cada intervalo (que comienza medio punto antes de su puntuación inferior y acaba medio punto después de la puntuación superior). Cuando se opera con este tipo de distribuciones, cada intervalo se representa por su marca de clase, X i , igual que la puntuación directa en el caso de datos no agrupados.

En el caso en que las frecuencias correspondieran a una distribución de frecuencias de variables cualitativas, como pueden ser los diferentes estados civiles o los grados universitarios, la representación se denomina diagrama de barras (figura 9).

## 2.2.- Las Medidas de Tendencia Central

Pues bien: cuando en Estadística se habla de representación de un conjunto de datos se piensa generalmente en las medidas denominadas de posición o tendencia central, alguna tan conocida como la media aritmética; junto a ella, la mediana y la moda.

Si en la vida ordinaria se dice de algo que está de moda estamos afirmando que es lo que más se lleva. Por ello, podemos representar los 100 valores anteriores por el que más se da, al que denominamos Moda (Mo) o, como otros dicen, Modo. Este valor es el 77, con los datos originales, o el 75,5 (marca de clase del intervalo con el mayor número de casos o frecuencia) en la distribución por intervalos.

Otro valor representativo es la Mediana (Md). Basta con ordenar de mayor a menor, o viceversa, la serie original y contar hasta encontrar el que ocupa el lugar central. Si la serie tiene un número par de casos, la Md será el valor medio de los dos centrales. En nuestro caso, con los datos originales, tales puntuaciones son iguales (76) por lo que la Md. coincide con ellos.

Ahora bien: si analizamos la situación, podemos ver que, en el primer caso, solo cuenta la puntuación que más se repite, mientras en el segundo la única que se toma en consideración es la que ocupa el lugar central, sin que ni siquiera importe cuál es su valor.

Son dos limitaciones a tener en cuenta. Ambas limitaciones son superadas por la más completa de estas medidas, la Media o Media aritmética, ya que todas y cada una de las puntuaciones de la serie contribuyen a configurarla en proporción a su valor. Por ello, para su cálculo no importa cuál sea la más repetida o cuál ocupe un determinado lugar en la serie ordenada; de hecho, no es preciso ordenar la serie sino sumar todas las puntuaciones y dividir la suma por el número de casos (N). Para el cálculo de la Media se aplica la ecuación siguiente:

$$Ecuación\ 3: Media = \frac{\sum_{i=1}^{N} X_i}{N}$$

La parte superior de la ecuación debe leerse como sigue:súmense todas las puntuaciones X desde la puntuación i a la puntuación N, esto es, desde la primera a la última.

En el supuesto de calcular la media en una distribución de frecuencias, la anterior ecuación se convierte en esta otra (ecuación 4), donde el valor X i no es una puntuación directa sino la marca de clase del intervalo:

$$Ecuación\ 4: Media = \frac{\sum_{i=1}^{N} X_i f_i}{N}$$

Compruebe el lector la pequeña distorsión que se da entre este valor, 73,02, el más exacto, y el obtenido en el caso de la distribución de 11 intervalos, donde la suma de los

productos de las marcas de clase por sus frecuencias arroja un valor muy próximo: 7230, con la cual la media es de 72,3. Puede comprobar estos datos en la tercera fila de la tabla 3 y en la última columna.

El tipo de medidas que se utiliza más comúnmente para representar a un grupo es el de tendencia central o posición y, dentro de estas, la media aritmética es la más completa; pero solo debe utilizarse con variables medidas con escalas de razón o cociente y de intervalo. En ocasiones, cuando los rangos de una variable ordinal se aproximan razonablemente a una escala de intervalo, también se suele utilizar la media aritmética.

### 2.2.1 - Medidas de Dispersión o Variabilidad

Ponga ahora atención el lector a estas dos series de datos ya presentados anteriormente:

5, 5, 5, 5, 5, 5, 5, 5, 5, 5

10,10, 10, 10, 10, 0, 0, 0, 0, 0

Si calculamos la Mediana, en ambos casos es la misma: 5 en la primera serie y (10 + 0) : 2 = 5 en la segunda. Y si lo hacemos con la Media, en ambos casos obtenemos una media de 5.

Sin embargo, a nadie se le oculta que estamos ante dos conjuntos de datos radicalmente diferentes, a pesar de que el valor representativo Media sea el mismo. Para hacer más realista el caso, piense en un profesor que tiene no 10 alumnos sino 20 o 30, en dos clases distintas: en la primera, los 20 o 30 niños, con puntuaciones de 5 en Matemáticas y en la segunda, con la mitad de casos con 10 y la otra mitad con 0. Parece obvio que no debería actuar del mismo modo en ambas clases.

Un tipo de medidas representativas diferente del anterior (medidas de posición o tendencia central) es el denominado de dispersión, que nos informa de esta característica.

Si en la primera de las dos series anteriores la dispersión en nula, dado que todas las puntuaciones coinciden con la Media, en el segundo es máxima ya que todos los casos se sitúan en los extremos.

En un caso como este, basta fijarnos en lo que se conoce como rango de la serie para hacernos una idea clara del grado de dispersión. Pero lo representado en ambas series no es lo habitual. Ni, por lo general, todos obtienen la misma puntuación ni se da una fractura tan grande entre los miembros del grupo.

Para apreciar la magnitud de la dispersión contamos con medidas específicas, tales como la desviación mediana, la desviación media, la desviación típica o la varianza.

El mismo nombre de la primera –desviación mediana- ya nos sugiere en qué consiste: es la media de las desviaciones de las puntuaciones con respecto a la Md del grupo. En el caso de la desviación media se trata, también, de la media de las desviaciones, pero ahora tomando como referencia la media aritmética.

Ahora bien: podemos comprobar qué es lo que pasa cuando hacemos estas operaciones en la siguiente serie, donde la media 5: (50 : 10) y Md es 6 (Figura 10).

Figura 10: Tratamiento de los Datos (X i) para el Cálculo de Medidas de Dispersión

| $X_i$ | 1 | 1 | 2 | 3 | 6 | 6 | 7 | 7 | 8 | 9 | ∑= 50 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $X_i$ – Md | -5 | -5 | -4 | -3 | 0 | 0 | 1 | 1 | 2 | 3 | -10 |
| $\|X_i$-Md$\|$ | 4 | 4 | 3 | 2 | 1 | 1 | 2 | 2 | 3 | 4 | 26 |
| $X_i$-Media | -4 | -4 | -3 | -2 | 1 | 1 | 2 | 2 | 3 | 4 | 0 |
| $(X_i$-Media$)^2$ | 16 | 16 | 9 | 4 | 1 | 1 | 4 | 4 | 9 | 16 | 80 |

Como se puede apreciar, en el primer caso obtenemos una suma positiva o negativa según que la distribución tienda a los valores inferiores o superiores a la Md (en este caso, los valores son negativos). Pero en el segundo la suma da, y siempre dará, 0 como consecuencia de las propiedades de esa medida de posición. Por eso, en el caso de la desviación mediana tendremos que tomar las desviaciones en valor absoluto (lo que se representa por el símbolo | |) y trabajar con la suma de las mismas.

$$\text{Ecuación 5: } DMd = \frac{|\sum_{i=1}^{N} X_i - Md|}{N}$$

DMd = 26 / 10 = 2,6

No obstante, no es esta la medida de dispersión más utilizada. Siempre que es posible, se acude a la desviación típica, representada por s, y a su cuadrado, conocido como varianza, representada por s 2 .

En ambos casos, las desviaciones con respecto a la Media (X i – Media) se elevan al cuadrado a fin de evitar que la suma dé 0. Pues bien: la varianza (s 2 ) es la media de las desviaciones de las puntuaciones individuales con respecto a la media, elevadas al cuadrado; por su parte, la desviación típica (s) es la raíz cuadrada de la anterior.

$$\text{Ecuación 6: } s = \sqrt{\frac{\sum_{i=1}^{N}(X_i - Media)^2}{N}} = \sqrt{\frac{80}{10}} = 2{,}828$$

$$\text{Ecuación 7: } s^2 : \frac{\sum_{i=1}^{N}(X_i - Media)^2}{N} = 8$$

Junto a las medidas de posición, podemos caracterizar un grupo con las de dispersión o variabilidad, que nos ofrecen una idea del grado de concentración de las puntuaciones directas en torno a la media, lo que tiene evidentes aplicaciones para la práctica profesional. Hemos citado, como fundamentales, la desviación media, la desviación típica* y la varianza.

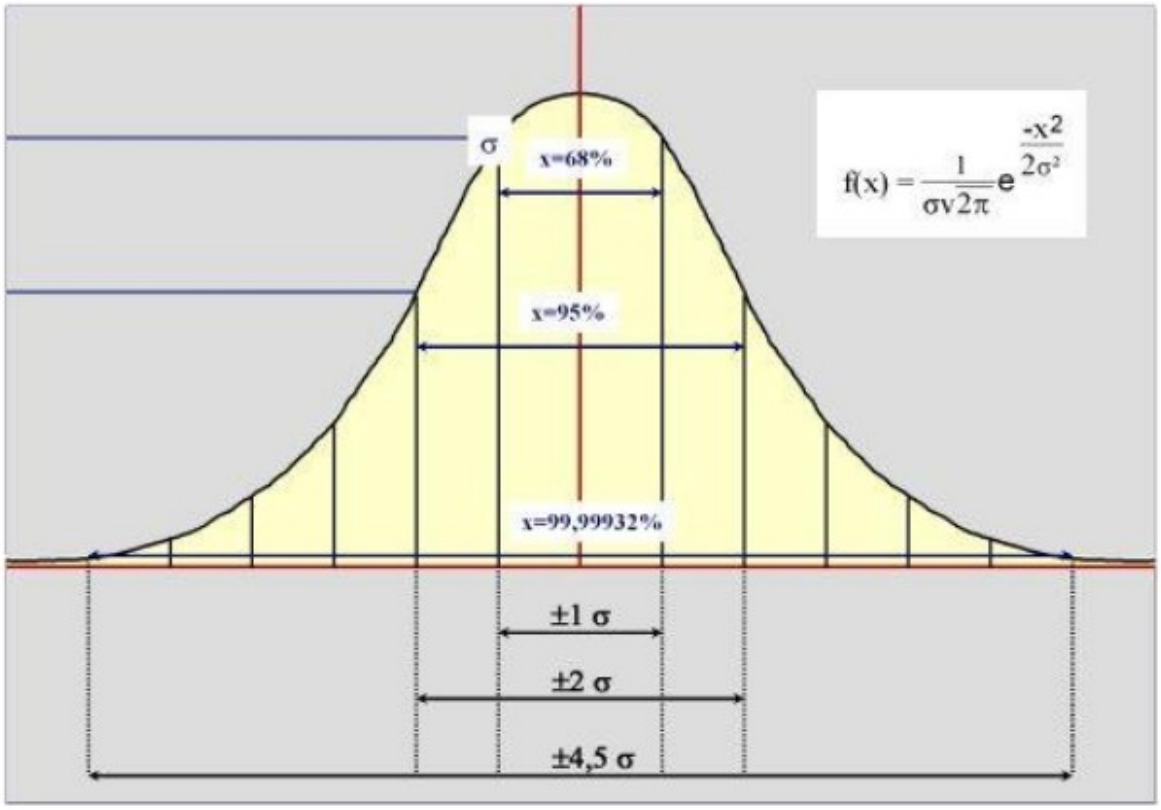
Estas medidas tienen su uso más frecuente en la denominada Estadística inferencial; una utilidad muy común e importante es la de interpretar una puntuación individual en el marco de una distribución normal (campana de Gauss) como veremos más adelante.

Suponiendo que nuestra distribución empírica de datos se acomoda al modelo normal podremos interpretar la puntuación de un sujeto cualquiera viendo cuántas unidades de s se aparta de la media del grupo, algo que podemos traducir fácilmente a porcentajes como tendremos ocasión de ver.

Esa puntuación individual, basada en s, se conoce como puntuación típica (z) a la que ya nos hemos referido, e indica en cuántas desviaciones típicas se aparta un sujeto de la media del grupo (Ecuación 2).

Aunque tendremos ocasión de verlo con más detalle, lo podemos apreciar en el siguiente gráfico de la curva normal de probabilidades (Figura 11):

Figura 11: Curva normal de probabilidades o Campana de Gauss




Cualquier puntuación individual (X i ) ocupa un lugar en la curva, por encima o por debajo de la ordenada de la Media (línea roja vertical), que la divide en dos partes simétricas. Las puntuaciones cercanas a la Media se encuentran a su derecha o a su izquierda, según sean mayores o menores que ella. Una puntuación X i que se aparte una desviación típica por encima o por debajo de la media se situará en la ordenada correspondiente del gráfico (± σ). Pero de esto hablaremos más adelante.

Baste decir ahora que la Estadística hace sus verdaderas aportaciones en lo que denominamos inferencia, que no es sino el proceso por el cual estimamos determinados valores de una variable en el conjunto total de casos (población) a partir de los medidos en una muestra de la misma. Los valores medidos en la muestra se denominan estadísticos y se representan como hemos hecho hasta ahora (M, Md, Mo, DMd, s, s 2 , ...) Los valores estimados en la población se denominan parámetros* y para ellos

utilizamos letras griegas (para el parámetro Media utilizamos, para la desviación típica, σ) .

Un ejemplo claro y sencillo: un profesor con 4500 puede tomar una muestra de los mismos de 150, obtener su media y estimar cuál será la media (μ) de los 4500. Y lo mismo con la desviación típica (σ) .

Otro: en las encuestas sobre intención de voto, se suelen tomar muestras de no más de 2 o 3 mil sujetos; a partir de sus respuestas se estima la intención de voto de los varios millones de chilenos que votarán.

Sin entrar en detalles, se comprende:

a) Que los datos fiables son los medidos en la muestra

b) Que los datos estimados en la población podrán apartarse en mayor o menor grado del verdadero valor.

c) Que la precisión de la estimación depende de la calidad de la muestra

d) Que los datos más útiles son los estimados a pesar del error de estimación que les afecte.

Cuanto más seguridad desee el investigador para sus estimaciones, más calidad deberá tener su muestra, esto es: más representativa de la población, lo que exige un tamaño suficiente y una selección imparcial de los sujetos, por lo general aleatoria. Para hacernos una idea de lo que entendemos por representatividad podemos acudir a una fotografía con respecto a la persona. Las fotografías pueden ser más o menos fieles al sujeto fotografíado.

Pues bien: para esos procesos de inferencia, las medidas de dispersión más utilizadas son la varianza y la desviación típica. Su cálculo es sencillo a partir de los datos de la tabla 4, ya que no es sino la media de las desviaciones elevadas al cuadrado, en el primer caso; en el segundo, es la raíz cuadrada de dicho valor. Cuestión diferente, como veremos, es la de su interpretación.

En nuestro caso, tal suma alcanza el valor de 80, por lo que la varianza será:

$$s^2 = \frac{(\sum(XI - Media)^2}{N} = 80 : 10 = 8,$$

$$\text{y la desviación típica} \quad s = \sqrt{\frac{80}{10}} = 2{,}828$$

Preciso es reconocer que no resultan de fácil comprensión ambos conceptos. Asumamos la idea de que se trata de la media de las desviaciones con respecto a la media (en el caso de la varianza), y de la raíz cuadrada de esta en el segundo.

Pero avancemos la importancia que tendrá la segunda cuando iniciemos el estudio de los modelos de probabilidad, como es la curva normal o campana de Gauss, de gran importancia y uso (la desviación típica) o las pruebas de significación estadística, como la

prueba F, para decidir si es razonable o no tomar en consideración determinadas diferencias (la varianza).

Un problema de estas medidas es su difícil interpretación; ni es fácil decidir sobre el grado de dispersión de una serie (si es poca, media o elevada) salvo si fuera nula, cuyo valor es 0, ni, mucho menos, decidir si una serie es más o menos dispersa que otra. A este último aspecto daremos respuesta mediante el coeficiente de variación.

Por el momento, dejémoslo ahí y avancemos con otras medidas de dispersión, como el recorrido semiintercuartílico y el coeficiente de variación.

Si la desviación típica* –s- se utiliza mucho en la estadística descriptiva, la varianza –s 2 – ofrece grandes aplicaciones en la inferencial.

Otras medidas a tener en cuenta son el recorrido intercuartílico –el que va entre los cuartiles 1 y 3- y el semiintercuartítico.

Ya conocemos la Mediana, medida de posición. Pues sepamos que la Md, que deja por encima y por debajo de sí al 50 % de los casos, equivale a lo que denominamos cuartil 2 (Q 2 = Md). Si cada una de las mitades se divide a su vez en partes iguales, la serie total queda dividida en cuatro partes mediante tres cuartiles: Q 1, Q 2, Q 3 . Pues bien, el 50 % central de la serie se denomina recorrido intercuartílico, y su división por 2 recorrido semi-intercuartílico.

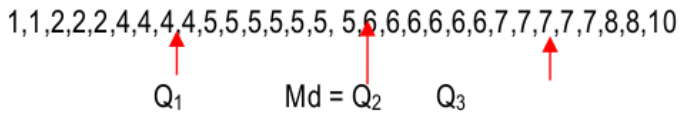
Su valor nos da información sobre la dispersión de la serie, como fácilmente se desprende de las tres siguientes series de datos: no es lo mismo que en una serie el 50 % central se encuentre ente puntuaciones muy próximas que el que para reunir ese 50 % tengamos que apartarnos ampliamente de la mediana del grupo. Veamos las tablas siguientes:

Figura 12: Distribución de frecuencias (f i ) y de frecuencias acumuladas (f a )

| $X_i$ | 1 | 2 | 4 | 5 | 6 | 7 | 8 | 10 | N |
|---|---|---|---|---|---|---|---|---|---|
| $f_i$ | 2 | 3 | 4 | 7 | 6 | 5 | 2 | 1 | 30 |
| $f_a$ | 2 | 5 | 9 | 16 | 22 | 27 | 29 | 30 | |

Sin entrar en detalles, la Md es 5; y los Q 1 y Q 3 4 y 7. Por tanto, el 50% de los casos se encuentra entre 4 y 7, siendo ese el valor de tal recorrido. Lo podemos apreciar fácilmente si la serie anterior la convertimos en datos originales, sin agrupar por frecuencias:

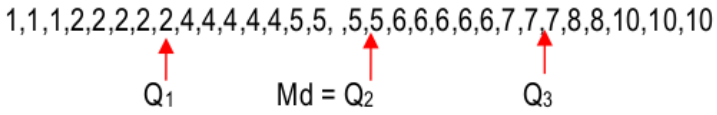
Figura 13: Distribución de frecuencias (f i ) y de frecuencias acumuladas (f a )




| $X_i$ | 1 | 2 | 4 | 5 | 6 | 7 | 8 | 10 | N |
|---|---|---|---|---|---|---|---|---|---|
| $f_i$ | 3 | 5 | 5 | 4 | 5 | 3 | 2 | 3 | 30 |
| $f_a$ | 3 | 8 | 13 | 17 | 22 | 25 | 27 | 30 | |

Aquí Q 1 y Q 3 son 2 y 6, respectivamente; por tanto, la serie presenta mayor dispersión*; es más plana que la anterior, que tiene mayor apuntamiento en los valores centrales. Veámoslo con datos sin agrupar:

Figura 14: Distribución de frecuencias (f i ) y de frecuencias acumuladas (f a )




| $X_i$ | 1 | 2 | 4 | 5 | 6 | 7 | 8 | 10 | N |
|---|---|---|---|---|---|---|---|---|---|
| $f_i$ | 8 | 6 | 3 | 2 | 2 | 1 | 6 | 2 | 30 |
| $f_a$ | 8 | 14 | 17 | 19 | 21 | 22 | 28 | 30 | |

## 2.3 - Las Medidas de Forma

En nuestro recorrido por las medidas de representación hemos visto las de posición o de tendencia central y las de dispersión*.

Utilizadas conjuntamente, tenemos una valiosa información para hacernos una idea de las características de un grupo. Pero podemos mejorar tal información mediante otras dos medidas de interés, no tanto por sus propias aportaciones como por lo que contribuyen a la caracterización del grupo; nos referimos a las de simetría y de apuntamiento, denominadas en algunos manuales como medidas de forma por ofrecer información sobre la forma general de la distribución de los datos.

Veamos estas series de datos (Series 3 a, b, c, d, e):

5, 5, 5, 5, 5, 5, 5, 5, 5, 5

1, 1, 1, 1, 1, 1, 1, 1, 1, 1

9, 9, 9, 9, 9, 9, 9, 9, 9, 9

1, 2, 4, 4, 5, 5, 6, 6, 8, 9

1, 1, 1, 3 5, 7, 8, 9, 9, 9

Las tres primeras series tienen la misma forma, una forma uniforme o plana; la diferencia se da en que los valores son medios en a) y extremos en b) y en c). La serie d) es más habitual: los valores extremos son menos frecuentes que los medios. Y la serie e) presenta una distribución menos frecuente, con más casos en los extremos que en el centro.

Si centramos nuestra atención en d) observamos que el valor más frecuente, el 5, está en el centro, y que tiene tantos valores a su izquierda como a su derecha; además, sus frecuencias descienden hacia ambos extremos en la misma forma: 2, 1 y 1 casos. Si representáramos la serie y la dobláramos por la mitad apreciaríamos su simetría.

Las medidas de forma nos ofrecen una idea de dos características del grupo como tal: el grado en que se acercan a la simetría, característica del modelo normal, y el de apuntamiento, más o menos equilibrado.

### 2.3.1 - Simetría/Asimetría

Pues bien; una medida de forma es la que nos indica su simetría o, mejor, el grado de asimetría de una distribución empírica; se representa por g 1 y mide el grado de asimetría de una serie de puntuaciones, esto es: la medida es que la serie empírica se aparta de una distribución simétrica, característica propia de las distribuciones denominadas normales, esto es, de las que siguen el modelo de la denominada curva normal de probabilidades* o campana de Gauss, una de cuyas características definitorias es la de ser simétrica con relación a la ordenada de la media.

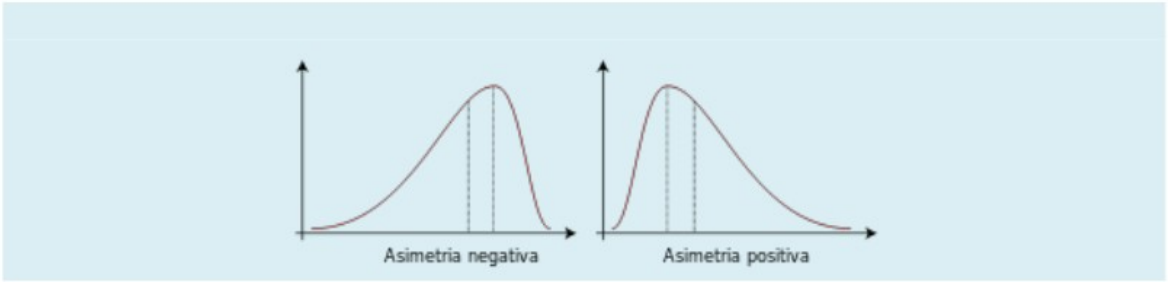
La medida del grado de asimetría, denomina coeficiente de asimetría, se representa por g 1 y se obtiene mediante la siguiente ecuación:

$$\text{Ecuación 8: } g_1 = \frac{\sum_{i=1}^{N}(xi - Media)^3}{Ns^3}$$

Cuando el número de valores de una distribución es mayor en la parte inferior a la media que en la superior a la misma, la distribución se muestra asimétrica hacia la izquierda, y hacia la derecha en caso contrario. En el primer caso g 1 < 0 y la asimetría se considera negativa; en el segundo, g 1 > 0, y la asimetría es positiva.

Si las diferencias entre los valores positivos y negativos en (X i – media) tienden a 0, la distribución se considera simétrica. La elevación de este valor al cubo se debe a que se trata de evitar que la ∑ (Xi-Media)=0, como nos ocurría en el caso de la varianza. En la figura 15 se presentan sendos ejemplos:

Figura 15: Distribuciones con Asimetría Positiva y Negativa




Las medidas de asimetría nos permiten calificar la distribución de las puntuaciones de un grupo como “normal” o como asimétricas, en mayor o menor grado, bien sea asimetría positiva o negativa.

### 2.3.2 - Apuntamiento o Curtosis

También con la serie d) podemos cuantificar su apuntamiento (simbolizado por g 2 ) esto es: el grado en que las puntuaciones centrales se concentran en torno a la media del grupo. El apuntamiento también recibe el nombre de curtosis.

Sin entrar en explicaciones que no vienen al caso, diremos que el apuntamiento normal se representa por g 2 = 3; valores de g 2 > 3 representan una distribución que recibe el nombre de leptocúrtica, mientras que en el caso de distribuciones con g 2 < 3, más achatadas, la distribución se denomina platicúrtica. La normal, obviamente, recibe el nombre de mesocúrtica.

La distribución leptocúrtica no solo tiene un mayor apuntamiento central sino que los valores extremos presentan, también, mayores frecuencias que en la normal. Por tanto, si un profesor está ante una distribución leptocúrtica sabe que sus alumnos se concentran más en el centro que en los extremos y que las puntuaciones extremas presentan frecuencias más elevadas que las que se darían si la distribución fuera normal.

El apuntamiento se obtiene mediante:

$$Ecuación\ 9: g_2 = \frac{\sum_{i=1}^{N}(X_i - Media)^4}{ns^4}$$

En la figura siguiente se pueden apreciar curvas con diferente grado de apuntamiento, superior e inferior al normal.

Figura 16: Distribuciones Leptocúrticas, Mesocúrtica y Platicúrticas

## 2.4 - Síntesis

**La Estadística y los números**

La Estadística* es una ciencia que trabaja con números. Los números se obtienen a partir de medir, pesar o contar objetos, sean estos directamente observables o no; por objeto entiéndase cualquier realidad que pueda medirse, pesarse o contarse, tanto si es animada como inanimada, si es persona o cosa, si es directamente accesible como si no.

La calidad de los números depende, fundamentalmente, de la posibilidad de aplicar a los objetos medidos unidades de medida fiables y válidas.

Frente a ciertas medidas referidas a objetos directamente accesibles, como la edad, el peso, la talla..., en nuestros ámbitos debemos acudir a "objetos" no directamente observables, lo que exige su definición (una definición denominada por lo general constructo) y la construcción de instrumentos adecuados para atribuirles valores. Para construirlos se procede a la denominada definición operativa de la variable.

Según sean los números obtenidos de las medidas de los objetos será o no lícito aplicarles ciertas propiedades y utilizarlos en determinadas operaciones matemáticas.

Los más perfectos pertenecen a las denominadas escalas de razón o cociente, seguidos por los de escala de intervalo, de las ordinales y, por último, de las nominales.

**Interpretación de los números**

Los números obtenidos a partir de la utilización de instrumentos esconden información que es preciso extraer.

Los números referidos a sujetos concretos (alumnos, pacientes, partidos políticos, desempleados...) no son fácilmente interpretables en sí mismos. Podemos hacerlo si

conocemos el “suelo” y el “techo” (puntuaciones mínima y máxima) del instrumento de recogida de datos y la unidad de medida, pero es más frecuente situar esas puntuaciones en el conjunto del grupo del que forma parte.

Entre las medidas individuales más habituales, podemos citar la de desviación (X i – Media aritmética = x i )*, la puntuación típica (z i )*, los cuantiles (cuartiles, deciles, percentiles), la Edad Mental (EM) o el cociente intelectual (EM / Edad cronológica).

Entre los procedimientos más sencillos para extraer información de los números se encuentra la simple ordenación de los mismos. Cuando el conjunto de valores de un grupo es elevado, la simple ordenación puede no ser suficiente para apreciar las características que lo definen.

Entonces podemos acudir a su reducción, haciendo que, sin alterar, o alterando de modo mínimo los datos originales, podamos hacernos una idea de las características del grupo con unos pocos valores.

Esta forma de actuar consiste en construir distribuciones de frecuencias en las que cada valor (X i ) va acompañado del número de veces que aparece (frecuencia: f i ).

Cuando estas distribuciones mantienen todo los valores originales, la distribución no se altera en absoluto. Sin embargo, en ocasiones, cuando la serie tiene un muy amplio recorrido (distancia entre los valores máximo y mínimo) puede ser conveniente que la distribución se reduzca construyendo intervalos de amplitud mayor que 1 (I i ) incluyendo para cada intervalo el número de casos –frecuencia- del conjunto de puntuaciones del intervalo.

En estos casos, la denominada marca de clase o valor medio del intervalo (X i ) se toma como representativa del intervalo a los efectos de los cálculos, lo que puede representar pequeñas desviaciones –positivas o negativas- entre los resultados de los cálculos con datos originales o de una distribución de esta naturaleza.

Estas desviaciones, por lo general, serán pequeñas porque, habitualmente, las diferencias positivas en unos casos se compensarán con las negativas en otros.

Teniendo en cuenta las limitaciones de los datos en fiabilidad y validez estas pequeñas desviaciones no deberían preocuparnos. La apariencia de de exactitud que nos da una calculadora con muchos decimales no refleja la realidad de los valores medidos, afectados por las limitaciones de los instrumentos de medida.

Caracterización y representación de grupos

La reducción de datos puede suponer una notable simplificación de los datos originales, haciéndolos más manejables; pero la Estadística nos permite algo más: representar el conjunto por medio de unas medidas que nos informan de las características más importantes del conjunto de datos.

Como toda representación nunca será tan perfecta con los datos originales, pero mientras estos, si son numerosos, se hacen muy difíciles de comprender y de tratar, aquellos los representan con la calidad suficiente para comprender la naturaleza y características del conjunto.

Tres son los tipos de medidas que nos ayudan a comprender las características de un grupo (ver cuadro siguiente):

Cuadro 1: Medidas Representativas de Grupo

| POSICIÓN O TENDENCIA CENTRAL | DISPERSIÓN* | FORMA | |
|---|---|---|---|
| | | SIMETRIA | APUNTAMIENTO O CURTOSIS |
| Moda: Mo | Recorrido | $g_1$ | $g_2$ |
| Mediana: Md | Desviación media | $g_1 > 0$. Asimetría positiva | $g_2 > 3$. Leptocúrtica |
| Media: M | Desviación típica: s | $g_1 = 0$. Simétrica Normal | $g_2 = 3$. Mesocúrtica |
| | Varianza: $s^2$ | $g_1 < 0$. Asimetría negativa | $g_2 < 3$. Platicúrtica |
| | Recorrido semi-intercuartílico | | |
| | Coeficiente de variación | | |

Las medidas de posición nos informan sobre la tendencia de la distribución de datos a acumularse en el centro de la misma (de ahí su otra denominación: de tendencia central) o Entre las medidas de posición, la más perfecta es la Media aritmética (por lo general denominada Media), dado que en ella influyen, de modo proporcional a su valor, todas y cada una de las puntuaciones de los datos originales. Resulta especialmente adecuada para medidas de razón o de intervalo.

Le mediana (Me o Md según los textos) también es una importante medida, pero tiene como inconveniente que en ella las puntuaciones no influyen por su valor sino por el lugar que ocupan, de modo que series muy diferentes pueden tener la misma mediana con solo mantener la misma puntuación central. Está especialmente adecuada a medidas de escala ordinal.

La Moda o Modo (Mo), poco utilizada, solo indica el valor más repetido. Se aplica fundamentalmente a puntuaciones de escala nominal.

Las medidas de dispersión son, probablemente, las más relevantes en el análisis de los datos numéricos, especialmente en la Estadística inferencial. Nos informan sobre el grado en que las puntuaciones se concentran o se separan de la media del grupo. En Estadística la dispersión de las puntuaciones es una cualidad o característica de gran valor y utilidad, como tendremos ocasión de ver.

Las más importantes son las más abstractas, en concreto la varianza (s 2 ) o media de las desviaciones de las puntuaciones con respecto a la media, elevadas al cuadrado, y la desviación típica o raíz cuadrada de la anterior.

En sí mismas nos ofrecen una información valiosa sobre la concentración o dispersión de las puntuaciones de una serie, si bien su interpretación no es fácil.

Además, estas dos medidas se utilizan mucho en la inferencia estadística, proceso por el cual estimamos los valores que se darán en la población (conjunto total de datos) a partir de los medidos en una muestra de la misma,

Los valores medidos se denominan estadísticos* (media, desviación típica, varianza...) y los estimados se denominan parámetros* (estos se representan mediante las correspondientes letras griegas: μ, σ, σ 2 ...).

Para estimar los parámetros tendremos que servirnos de los modelos estadísticos y de la teoría de la probabilidad*. De este modo, cualquier valor estimado vendrá acompañado de la probabilidad* de que ocurra.

El modelo de referencia más habitual es el denominado normal. Tomándolo como referencia, decidimos si la distribución empírica es platicúrtica o leptocúrtica, si su asimetría es negativa o positiva.

Contamos con pruebas que nos permiten decidir si una distribución empírica se acomoda o no a la normal; en caso positivo, podemos aplicar a los datos empíricos las propiedades del modelo, pensando que las desviaciones apreciadas se deben a pequeñas imperfecciones en la selección de los datos.

En este proceder no hacemos sino algo habitual: nadie ha visto en la Naturaleza un cono, pero sí montañas más o menos cónicas (pensemos en el Teide). Pues bien: dando por bueno que el Teide no se aparta mucho de un cono ideal, podemos calcular, aproximadamente, su superficie y su volumen, aplicándole la fórmula del modelo, del cono.

Como se puede comprender, el problema es decidir si el objeto empírico se acomoda razonablemente al modelo; la Estadística nos ayudará a ello mediante pruebas denominadas de bondad de ajuste* (por ejemplo, para el caso del ajuste a la curva normal, la de χ 2 ; léase ji o chi cuadrado)

Las medidas de forma, como su nombre indica, nos ofrecen una visión global sobre la forma de la distribución, fijándose en dos aspectos fundamentales: la simetría y el apuntamiento. Para valorar tales características se toma como referencia la denominada distribución normal, que es simétrica respecto de la ordenada de la media y que tiene un apuntamiento normal –mesocúrtica- en sus valores centrales.

La asimetría puede ser negativa, cuando el valor del correspondiente coeficiente tiene valores negativos, quedando sesgada hacia la izquierda, o positiva, cuando el correspondiente valor es positivo, quedando sesgada hacia la derecha.

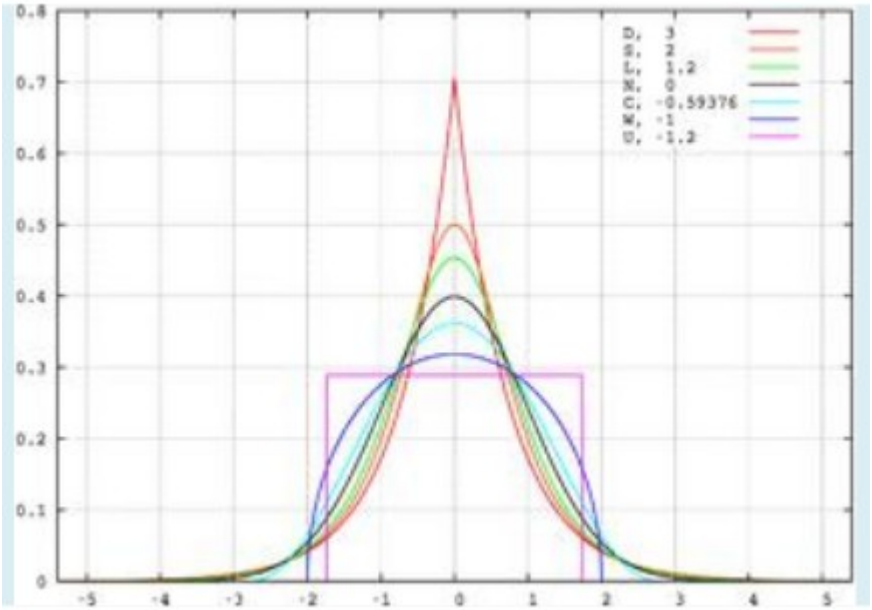
El apuntamiento normal nos sitúa ante distribuciones mesocúrticas, siendo leptocúrticas cuando el apuntamiento es mayor y platicúrticas si es menor.

Si se ofrecen datos de estos tres tipos de medidas, la caracterización de una distribución de puntuaciones es muy completa y, sobre permitirnos una comprensión profunda de sus características, nos facilitará la realización de determinados procesos de inferencia, entre los que destacamos, precisamente, la estimación de parámetros, con determinada probabilidad, y la realización de contrastes, mediante pruebas estadísticas que nos

permitirán tomar decisiones sobre los efectos de las variables independientes sobre las dependientes en los experimentos.

Por otra parte, los seres humanos estamos más habituados a comprender los fenómenos que ocurren ante nuestros ojos o que somos capaces de representar de forma intuitiva.

Pues bien: los números también pueden representarse mediante una serie de representaciones, que vamos a ver, y que nos facilitan la interpretación de forma más fácil; digamos, no obstante, que los números, unidos a sus representaciones gráficas, se complementan: estas ofrecen la visión intuitiva; aquello, la precisión.




# CAPÍTULO TRES

# Probabilidades

El siguiente capítulo busca introducir los conceptos básicos en el estudio de las probabilidades, además de visualizar su posible aplicación mediante ejemplos teóricos. Lo anterior con un enfoque en las Ciencias Sociales y tomando como base los aprendizajes básicos de la estadística para ocuparlos en el entendimiento del concepto de Probabilidad.

Concluido el capítulo se espera que usted logre los siguientes objetivos:

- Definir el concepto de Probabilidad
- Definir los conceptos de Experimento, Resultado y Evento.
- Determinar el Espacio Muestral que enmarca el estudio de un hecho determinado.
- Identificar el tipo de suceso al que se enfrenta al momento de abordar un hecho en particular.
- Comprender y diferenciar un hecho aleatorio de uno determinístico
- Ser capaz de calcular una probabilidad según el método más apropiado, poder entender que subyace ese método.
- Trabajar algebraicamente el cálculo de una probabilidad
- Comprender los conceptos de Probabilidad condicional e independiente y ser capaz de calcularlas
- Comprender el Teorema de la probabilidad total
- Calcular una probabilidad mediante el Teorema de Bayes

## 3.1 – Introducción

El estudio de las probabilidades se basa en la generación de una conexión entre hechos que han ocurrido y otros que pueden hacerlo, esto equivale a decir entre hechos determinísticos y aleatorios. Por ejemplo, cuando lanzamos una moneda al aire sabemos dos cosas: que la moneda caerá y que saldrá una de las caras de la monera, el segundo hecho es algo que no manejamos con certidumbre, antes de lanzar la moneda nadie podría aventurarse a decir con certeza que cara saldrá, este es un hecho aleatorio.

La importancia de la **probabilidad** para las ciencias sociales viene dado por lo anterior, poder realizar un nexo entre un hecho cierto, determinado, y algún/os hechos aleatorios nos puede ayudar a cuantificar la influencias o efectos. Mediante estos efectos podemos, no solo realizar una evaluación y análisis de los hechos que estudiamos, sino también obtener conclusiones capaces de ser generalizadas a eventos que no suceden controladamente o, inclusive, aun no suceden.

Pensemos en la educación, el hecho de que un alumno obtenga una nota es algo dado. Juan llega a su casa y muestra que obtuvo un 5.3 en su prueba de probabilidades. Su madre sabe que él estudio toda la semana, que asistió a todas las clases de probabilidades, que completó las guías de ejercicios y que su cuaderno está al día con toda la materia, pero no sabe, quizás lo intuye, la facilidad que tiene Juan para las probabilidades. Esto último es un hecho aleatorio y la pregunta es cómo se relaciona con su nota y si podemos realizar algo con esta información.

El ejemplo anterior hace necesario que conceptualicemos una serie de elementos necesarios para el estudio de las probabilidades y su mejor entendimiento. Partamos diciendo que una probabilidad es un número entre 0 y 1, este número puede ser multiplicado por alguna potencia de 10 para entregar un valor entero, ejemplo de esto es obtener una probabilidad de 0.7, multiplicarla por 100 y obtener un 70%.

***Probabilidad**: Valor entre 1 y 0, que puede ser multiplicado por alguna potencia de 10. Este valor nos habla sobre la posibilidad con que esperamos que un evento ocurra.*

Para poder obtener una probabilidad necesitamos que un evento se realice, generalmente diremos que necesitamos un experimento y su resultado. Entenderemos por **experimento** un hecho en particular, que buscamos cuantificar, y que aislamos de otras observaciones, esto implica que sólo ocuparemos una de las posibles observaciones, o hechos, que tengamos a disposición. Un **resultado** se entenderá como la realización de un experimento, tomando en consideración las consecuencias que tuvo este en las observaciones que se tomaron. Finalmente tendremos que un **evento** se entenderá como un conjunto de experimentos, así el evento será el hecho en sí.

***Experimento:** Realización de una observación en particular, buscando cuantificar las consecuencias de esta.*

***Resultado:** Realización de un experimento, son las consecuencias que este tuvo en la observación.*

***Evento:** Conjunto de Experimentos que se realizan, agrupan a todas las posibles observaciones de un hecho.*

Para aterrizar más los conceptos tomemos en consideración el ejemplo del lanzamiento de una moneda al aire. El hecho que ocurrirá será lanzar la moneda al aire, y las posibles observaciones serán los lados de la moneda, esto implica que el hecho tiene dos observaciones. El **experimento** será cuantificar cuantas veces sale el lado *cara* de la moneda si lanzo diez veces la moneda al aire, o cuantas veces sale el sado *sello* al lanzar la moneda al aire la misma cantidad de veces. Esto implica que tendremos dos experimentos frente al hecho. El **resultado** sería cuantas veces sale *cara* o *sello* al lanzar la moneda 10 veces al aire, recordemos que esto es la realización del experimento por lo tanto debe ser la cuantificación del mismo, por esto decimos que son las consecuencias. El **evento** sería *el lanzamiento de una moneda al aire 10 veces*.

El ejemplo anterior nos muestra que la probabilidad resultante de esto nos permite una aproximación hacia eventos futuros, no es necesario volver a realizar el experimento para saber la probabilidad de que salga *cara* al lanzar 10 veces una moneda al aire. Más aún, si repito este experimento en otra parte del mundo, los resultados debiesen ser homologables, por lo tanto es totalmente generalizable el resultado. Veremos más adelante que para que lo anterior sea totalmente cierto es necesario repetir el experimento muchas veces y así lograr que la probabilidad se acerque hacía su valor real.

El capítulo se desarrollará de la siguiente manera, a continuación se desarrollará el concepto de espacio muestral, aleatoriedad y sucesos que nos permitirán abordar de mejor manera el cálculo de una probabilidad. Luego veremos los distintos enfoques para realizar el cálculo de una probabilidad y, posteriormente, veremos el álgebra detrás de esos conceptos, ahí introduciremos la noción de probabilidad condicional y probabilidad independiente. Terminaremos el capítulo revisando el teorema de la probabilidad total y el teorema de bayes para el cálculo de probabilidades.

## 3.2 - Teoría de las Probabilidades

En el siguiente apartado revisaremos conceptos fundamentales para entender el cálculo y el trabajo con las probabilidades. Estos conceptos se enmarcan dentro de la teoría básica que subyace el estudio de la probabilidad y nos servirán para situar el contexto en el cual nos estamos moviendo. Los conceptos fundamentales que revisaremos serán aleatoriedad, sucesos y espacio muestral.

### 3.2.1 – Aleatoriedad

La aleatoriedad se define como un proceso cuyo resultado no es previsible con certeza, por lo tanto se asocia a que este tiene relación con el azar. Esta definición se puede aplicar a la estadística mediante el término *experimento aleatorio*, el cual une dos conceptos, primero el término experimento y luego el concepto de aleatoriedad.

El experimento aleatorio hará referencia a una realización en particular cuyos resultados no tienen certidumbre, y por lo tanto interviene el azar. Esto implica que bajo las mismas condiciones iniciales podemos tener un número variado de resultados. Como contraposición existen los ***fenómenos deterministas*** los cuales son hechos realizados intencionadamente para lograr un resultado, y consecuencias, en particular.

Generalmente en las ciencias sociales intentamos trabajar bajo la aleatoriedad, en específico buscando hechos que puedan homologarse a experimentos aleatorios. Esto

debido a que permitimos que el azar o la naturaleza actúe bajo sus reglas con lo cual logramos tener una mejor impresión sobre los resultados de un hecho en particular.

Para que un experimento se diga aleatorio deben verificarse las siguientes condiciones:

- El conjunto de todos los posibles resultados es conocido con anterioridad. Esto implica que podemos establecer el espacio muestral del experimento, sea discreto o continuo.
- Si el experimentos e repite, el resultado no es conocido con certeza, sólo es posible aproximarnos mediante el cálculo de probabilidades.
- El experimento puede ser repetido bajo las mismas condiciones iniciales, esto implica que el experimento tiene un carácter de genérico.

### 3.2.2 – Sucesos

Hace referencia a un conjunto de resultados posibles para un experimento dado, esto implica que un suceso en particular será una consecuencia de un experimento que realicemos. Diremos que un suceso ocurre si alguno de los resultados que contiene se da.

Podemos encontrar nueve tipos de sucesos:

- ***Suceso elemental***: Cada uno de los resultados posibles, que conforman el espacio muestral, frente a un experimento cualquiera. Ejemplo, que al lanzar un dado salga el número cinco.
- ***Suceso compuesto***: Es un subconjunto de los posibles resultados esperados. Ejemplo, que al tirar un dado saliera un número impar o un múltiplo de dos.
- ***Suceso seguro***: Está formado por cada uno de los posibles resultados. Ejemplo, lanzar dos dados y obtener una suma menor a 13.
- ***Suceso imposible***: Es un conjunto que no contiene ningún resultado posible frente a un experimento cualquiera. Ejemplo, lanzar un dado y obtener un número mayor a diez.
- ***Sucesos compatibles***: Dos sucesos son compatibles cuando alguno de los elementos que contienen (algún suceso elemental) se repite, o sea tienen a lo menos un suceso en común.
- ***Sucesos incompatibles***: Dos sucesos son incompatibles cuando no poseen ningún suceso elemental en común.
- ***Sucesos independientes***: Dos sucesos son independientes cuando la probabilidad de que uno se dé, o suceda, no se ve afectada por la probabilidad de que el otro suceso se realice.
- ***Suceso dependiente***: Dos sucesos son dependientes cuando la probabilidad de realización de uno se ve afectada por la de otro.
- ***Suceso contrario***: Se define como el suceso que ocurre cuando otro no ocurre. Para denotarlo se dice "*el suceso contrario a A*" y hace referencia al suceso que ocurre cuando A no ocurre.

### 3.2.3 - Espacio Muestral

Un espacio muestral o espacio de muestreo consiste en los posibles resultados que tiene un experimento determinado. Esto implica conocer las consecuencias que tendrá el experimento que ejecutaremos, y a la vez, ser capaces de empezar a diferenciar los posibles resultados que podríamos obtener.

Un ejemplo de esto es el experimento de lanzar dos monedas al aire, acá tenemos que por cada lanzamiento tendremos dos monedas y en cada moneda tenemos dos opciones, *cara* o *sello*. Esto nos hace pensar que nuestros posibles resultados serán pares, debido a que cada experimento será lanzar dos monedas, estos se pueden resumir en el siguiente conjunto: *{cara,cara}, {cara,sello}, {sello,cara}, {sello,sello}*. Vemos que el espacio muestral debe contener todos los resultados posibles frente a un experimento cualquiera.

El espacio muestral nos permite saber cuales son los resultados posibles frente a un experimento cualquiera que deseemos realizar. Lo relevante es que identificar el espacio muestral nos permitirá determinar si los resultados esperados nos sirven para los objetivos que buscamos. Quizás sea necesario re evaluar la pregunta de interés, buscar otro experimento o cambiar el hecho en investigación, para lograr el objetivo que buscamos.

Tenemos dos grandes tipos de espacios muéstrales: *discretos* y *continuos*. La diferencia entre ambos hace referencia el tipo de sucesos que contienen, específicamente si estos son finitos o no.

#### 3.2.3.1- Espacio muestral Discreto

El espacio muestral discreto es aquel en que los sucesos que forman parte de él son finitos o infinito numerables, esto implica que podemos decir que cuantos son de manera certera. Los elementos dentro del espacio muestral deben poder “contarse”, esto no implica que al definir el espacio podamos encasillar los elementos de manera finita, sino que los elementos que pertenezcan al espacio muestran sean finitos.

Dentro de esta noción encontramos cuatro categorías que nos permiten profundizar más en la conformación del espacio muestral, estas son:

- ***Espacio probabilístico discreto equiprobable***: El espacio muestral es finito de tamaño *n* y todo suceso tiene una probabilidad *E*. Esto implica que todo suceso tiene la misma probabilidad de ocurrencia que es igual a $\frac{1}{n}$ .
- ***Espacio probabilístico finito***: El espacio muestral es discreto finito, esto equivale a que podemos contarlo y sabemos que tienen un final cuantificable, además al menos dos sucesos cumplen con que su probabilidad de ocurrencia es distinta.
- ***Procesos estocásticos finitos***: Esto hace referencia a una sucesión finita de experimentos aleatorios, cada uno de los cuales tiene un número finito de resultados. La representación de estos procesos es mediante un diagrama de árbol.

- ***Espacio probabilístico infinito contable***: Es aquel espacio muestral en que sus elementos son discretos, se pueden contar, pero son infinitos. El mejor ejemplo es la probabilidad de que salga cara en un lanzamiento de la moneda al aire, esta probabilidad será de ½ pero también de la familia de múltiplos que de esa fracción nazca.

3.2.3.2- Espacio muestral Continuo

Este espacio muestral se caracteriza porque sus elementos, los sucesos que contiene, son infinitos incontables, vale decir, el espacio muestral será un conjunto infinito de resultados posibles. Generalmente intentaremos encasillarlo mediante reglas que sigan la teoría de conjuntos, pero sabremos que, antes de realizar el experimento, es difícil aventurar un posible resultado de forma certera.

En este tipo de espacio muestral tendremos dos categorías posibles, las cuales son:

- ***Espacio probabilístico continuo***: Es un espacio muestral con sucesos infinitos no numerables, esto implica que no es posible visualizar puntos concretos en el espacio o plano cartesiano. Por lo mismo trabajamos con intervalos observados y no con puntos u observaciones propiamente tal, esto implica que trabajaremos sobre intervalos como unidad básica de medida.
- ***Particiones***: Es un espacio muestral con sucesos infinitos incontables pero que se pueden agrupar en un conjunto numerable. Esto implica que podemos decir los elementos que conforman el espacio muestral y establecer reglas que hacen que un suceso sea parte y otro no.

## 3.3 - Enfoques de cálculo de probabilidades

Para poder calcular una probabilidad tendremos dos enfoques: el enfoque *objetivo* y el *subjetivo*. Dentro del primero encontraremos la ***probabilidad clásica*** y la ***probabilidad empírica***.

### 3.3.1 - Probabilidad clásica

Este enfoque parte del supuesto de que *todos los resultados posibles del experimento son igualmente posibles*, esto implica pensar que al momento de realizar un experimento todos los resultados que se pudiesen esperar tienen la misma posibilidad de ocurrencia. Dado esto el cálculo de la probabilidad viene dado por la división entre el número de resultados favorables y el número total de resultados.

$$Probabilidad\ Clásica = \frac{Número\ de\ resultados\ existosos}{Número\ de\ resultados\ totales}$$

Es necesario precisar que el número de resultados exitoso hace referencia al número de veces que se da el resultado que estamos “testeando”, en el ejemplo del lanzamiento de la moneda si queremos experimentar cuantas veces sale *sello* debiésemos dividir el número de veces que sale *sello* al lanzar la moneda sobre el número de resultados posibles. Lo anterior se ejemplifica a continuación:

***¿Cuál es la probabilidad de que salga sello al lanzar una moneda al aire?***

*Al lanzar una moneda al aire tenemos dos resultados posibles, que salga cara o que salga sello, esto implica que tenemos un resultado exitoso sobre dos posibles resultados.*

$$Probabilidad\ de\ que\ salga\ sello\ al\ lanzar\ una\ moneda\ al\ aire = \frac{1}{2}$$

Según la teoría clásica si los evento son mutuamente excluyentes y el conjunto de eventos es colectivamente exhaustivo, la suma de las probabilidades debe dar 1. Que un evento sea **mutuamente excluyente** significa que la realización de una observación implica la no realización de las otras, en el caso del ejemplo del lanzamiento de la moneda no puede salir *cara* y *sello* al mismo tiempo, esto implica que es un evento **mutuamente excluyente**.

Decir que el conjunto de eventos es **colectivamente exhaustivo** implica que al realizar un experimento siempre se debe dar, a lo menos, un evento de los experimentados, esto implica que al realizar el experimento debe siempre salir un resultado acorde a lo que intentamos medir, en el ejemplo de la moneda siempre al lanzarla hacia arriba saldrá *cara* o *sello*, jamás tendremos un resultado distinto, esto es **colectivamente exhaustivo**.

### 3.3.2 - Probabilidad empírica

Este método también se conoce como **frecuencia relativa** y basa en que saber cuántas veces se da un resultado sobre la cantidad total de intentos que se han efectuado en el experimento. Esto implica que la probabilidad nos da nociones sobre cuantas veces esperamos que suceda un resultado dentro de un evento.

$$Probabilidad\ Empírica = \frac{Número\ de\ veces\ que\ el\ evento\ ocurre}{Número\ total\ de\ observaciones}$$

Por lo tanto una probabilidad dependerá de cuantas veces se da un resultado en particular dentro de la experimentación que realizamos, esto produce que la probabilidad pueda variar a medida que aumentamos el número de veces que repetimos el experimento. Este efecto se conoce como ***La ley de los Grandes Números*** y hace referencia a que a medida que repitamos el experimento, la probabilidad de acercar a su valor real. Esto se aprecia en la siguiente figura.

| Número de | Veces que | Veces que | Probabilidad | Probabilidad |
|---|---|---|---|---|

| observaciones | sale cara | sale sello | relativa de las caras | relativa de los sellos |
|---|---|---|---|---|
| 10 | 3 | 7 | 0.3 | 0.7 |
| 100 | 40 | 60 | 0.40 | 0.60 |
| 1000 | 485 | 515 | 0.485 | 0.515 |
| 10000 | 4999 | 5001 | 0.4999 | 0.5001 |

Esto nos muestra que a medida que aumentamos el número de observaciones totales, la probabilidad relativa de los eventos se acerca a su valor real. Esto nos presenta algunas complicaciones, debido a que frente a muestras “pequeñas” debemos tener cuidado al calcular probabilidades relativas o por lo menos tener precaución al momento de generalizar resultados a partir de estas.

### 3.3.3 - Probabilidad subjetiva

Esta probabilidad se basa en el cálculo ocupando información que este en el ambiente sobre un hecho puntual. Lo anterior implica, muchas veces, la imposibilidad de realizar un experimento en torno al hecho que se quiere analizar, por lo tanto se ocupa la información disponible. Algunos ejemplos pueden ser:

- Probabilidad de que un equipo deportivo gane el título el próximo año
- Probabilidad de que usted contraiga matrimonio en los próximos 15 años
- Probabilidad de que Chile aumente su matriz energética no contaminante

Estas probabilidades se basan en hechos que no son posibles de experimentar, pero si es posible ocupar información existente, como número de matrimonios de personas según rango etario y condición socioeconómica, para calcularlas. Esto significa aproximarnos a la probabilidad con la finalidad de tener una “certeza” frente hechos que ocurrirán en el futuro y no podemos cuantificar en la actualidad mediante experimentación.

## 3.4- Reglas para calcular probabilidades

En este ocuparemos las reglas de *adición* y *multiplicación* para el cálculo de la probabilidad de dos o más eventos que se den simultáneamente.

### 3.4.1- Reglas de adición

Dentro de las reglas de la adición encontramos dos especificaciones que ocupan a la adición como eje fundamental del cálculo de dos o más eventos que se den de manera combinada, estás son la *regla especial de la adición* y la *regla general de la adición*.

#### 3.4.1.1- Regla especial de la adición

Para la aplicación de esta regla los eventos deben ser ***mutuamente excluyentes***, recordemos que esto implica que al ocurrir un evento ningún otro puede ocurrir. Esto implica que ocuparemos el conector lógico “*o*” e intentamos decir “*La probabilidad de que suceda el evento 1 o el evento 2*”, por lo tanto sumaremos ambas probabilidad para

obtener este resultado. Lo anterior es la llamada **regla especial de la adición** y se puede definir de la siguiente manera:

$$P(A\,o\,B)=P(A)+P(B)$$

Siendo *P(A o B)* la probabilidad de que suceda el evento A o el evento B, *P(A)* es la probabilidad de que suceda el evento A y *P(B)* la probabilidad de que suceda el evento B. Esto nos permite realizar una generalización de esta regla hacia infinitos eventos de la siguiente manera:

$$Regla\,especial\,de\,la\,adición:$$

$$P(A\,o\,B\,o\,C\,o\,D\,o\ldots\ldots\ldots o\,N)=P(A)+P(B)+P(C)+P(C)+P(D)+\ldots\ldots\ldots+P(N)$$

Dentro de este apartado cabe el estudio de las probabilidades complementarias, y por consiguiente la regla que emerge de ellas. Como su nombre lo dice, la ***regla del complemento*** hace referencia a cómo obtener la probabilidad del suceso complementario, vale decir la probabilidad de que el suceso no se realice. Esta probabilidad se basa en la siguiente igualdad

$$P(A)+P(A_C)=1$$

Donde P(A) es la probabilidad de que se realice el suceso A y P(Ac) es la probabilidad de que el complemento del suceso A se realice. Como vemos debemos despejar esta igualdad para obtener el resultado de la probabilidad complementaria al suceso A.

$$P(A_C)=1-P(A)$$

De esta forma obtener la probabilidad complementaria al suceso que estudiamos y obtenemos más información sobre el hecho planteado. La obtención de esta probabilidad no es trivial, puede ser más sencillo obtener la probabilidad de un suceso haciendo que este sea un complemento a indagar el suceso en sí, esto debido a los costos asociados que pueden existir en la experimentación.

### 3.4.1.2- Regla general de la adición

La característica distintiva de esta regla es que permite hacerlo para eventos o sucesos que no son mutuamente excluyentes, como visitar regiones de un país. La probabilidad que intentamos obtener es una que mezcle dos eventos que no son mutuamente excluyentes como puede ser "*La probabilidad de que juan haya estado en la primera región o en la décima región*".

El ejemplo ilustra que ambos sucesos pueden ser reales, pero yo quiero saber si es uno u otro, no ambos a la vez. Lo que hacemos para determinar la probabilidad exacta es sumar la probabilidad de que un suceso se realice a la del que se realice el otro suceso, pero restamos la probabilidad de que ambos se den a la vez. Restar la probabilidad de que

Juan haya ido a ambas regiones implica sacar de mi grupo de interés a todas las personas que cumplan con ambos sucesos a la vez. La fórmula de la regla se aprecia a continuación:

$$P(A\,o\,B)=P(A)+P(B)-P(A\,y\,B)$$

Para generalizar la fórmula basta con imaginar que queremos saber la probabilidad de que se den más hechos y tener cuidado en que debemos agregar todas las probabilidad conjuntas, vale decir, cuando se realizan dos o más suceso a la vez.

La diferencia entre ambas reglas está en la probabilidad de que dos, o más sucesos, se den a la vez. Esta probabilidad solo existirá (tendrá un valor mayor que cero) para el caso en que los eventos no sean mutuamente excluyentes, por lo tanto esta característica definirá que tipo de regla debo ocupar.

### 3.4.2- Reglas de multiplicación

La regla de la multiplicación nos permite calcular la probabilidad de que dos, o más sucesos, ocurran simultáneamente. El valor de esto es que podemos saber cuál es la probabilidad de que dos sucesos de interés se den simultáneamente en una persona, por ejemplo: la probabilidad de que Juan haya votado en las elecciones municipales y haya votado en las elecciones presidenciales. Tenemos dos hechos, una persona y una probabilidad que deseamos calcular. Tendremos dos reglas para la obtención de esta probabilidad, la *regla especial* y la *regla general*.

#### 3.4.2.1- Regla especial de la multiplicación

Esta regla impone la necesidad de que los dos sucesos sean **independientes**, esto implica que la realización de uno no afecta la probabilidad de realización del otro. Esto nos dice que la realización de un evento es independiente frente a la realización de otro evento que estemos analizando.

La independencia nos asegura que la probabilidad de que ambos eventos se realicen al simultáneamente será la multiplicación de la probabilidad de los eventos por separado, la **regla especial de la multiplicación** se muestra en la siguiente fórmula:

$$P(A\,y\,B)=P(A)\,x\,P(B)$$

Si deseamos generalizar la formula a más sucesos, basta con exigirle independencia entre ellos y multiplicar como se mostró en la formula. Que se le exija independencia a cada uno de ellos implica que ningún suceso puede afectar la probabilidad de ocurrencia de algún otro que deseemos analizar.

#### 3.4.2.2- Regla general de la multiplicación

Cuando dos sucesos no son independientes, vale decir la realización de uno si afecta la probabilidad de suceso del otro, diremos que los eventos son **dependientes**. Un ejemplo de este tipo de sucesos es sacar bolitas de bingo de una tómbola, el primer suceso será sacar una bolita par en el primer intento, el segundo será sacar una bolita par en el

segundo, dado que la cantidad de bolitas no será la misma en el segundo intento y quizás tampoco sea la misma cantidad de bolitas par, la probabilidad cambiará.

Esto muestra que la probabilidad del segundo suceso está condicionada por la probabilidad de realización del primero, esto genera un concepto conocido como **probabilidad condicional** que se verá en el siguiente apartado. La probabilidad condicional se definirá como la probabilidad de que ocurra un suceso, dado que ya ocurrió otro. Esta probabilidad es fundamental para el cálculo de la probabilidad mediante la regla de la multiplicación, la fórmula de esta es la siguiente:

$$P(A\ y\ B) = P(A)\ x\ P(B \vee A)$$

La probabilidad de que dos sucesos pasen simultáneamente será la multiplicación entre la probabilidad de que ocurra el primer suceso y la probabilidad condicional de que ocurra el segundo.

La generalización de esta regla no es trivial y se realiza condicionando la probabilidad de un evento por la realización conjunta de los otros, veamos cómo será la fórmula para tres eventos:

$$P(A\ y\ B\ y\ C) = P(A)\,P(B \vee A)\,P(C \vee A\ y\ B)$$

Esta generalización nos permite el cálculo de más sucesos que ocurran simultáneamente, lo que puede ser de bastante interés para las ciencias sociales. Generalmente intentamos tratar los sucesos por separado, pero mediante las reglas de la multiplicación hemos verificado que es posible el cálculo de probabilidades de hechos que suceden simultáneamente, solo debemos tener cuidado en diferenciar la independencia de estos sucesos.

## 3.5- Probabilidad condicional e independiente y teorema de la probabilidad total

Este apartado abordara tres tópicos importantes dentro del estudio de la probabilidad, *la probabilidad condicional*, *probabilidad independiente* y el *teorema de la probabilidad total*. Su estudio se realizará de forma conjunta debido a la conexión que existe entre ellos, los dos primeros son tipos de probabilidad que dependerán de la independencia que exista entre los sucesos que analizamos, por lo tanto ocuparemos las reglas de la multiplicación para poder calcularlo, y el último tópico ocupara los dos primeros para establecer un primer teorema acerca de la probabilidad.

### 3.5.1- Probabilidad condicional e independiente

Como se definió en el apartado sobre las reglas de la multiplicación, en particular sobre la regla general de la multiplicación, la **probabilidad condicional** es la probabilidad de que un evento suceda dado que ya sucedió otro evento. Esto quiere decir que la realización de un suceso estaría condicionada por la realización de otro, esto se conoce como dependencia.

La probabilidad condicional se entenderá como "*la probabilidad de que el evento A suceda, dado que sucedió el evento B*". Esto implica que nos interesa la fracción de eventos que ya se realizaron pero que influyen en la probabilidad de realización de otro, ocupando la notación anterior, no interesa saber que fracción dentro del conjunto de evento B ejerce una influencia en la realización de eventos A. La notación y fórmula de la probabilidad condicional son las siguientes:

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

Vemos que en numerador esta la probabilidad de encontrar elementos comunes en los eventos A y B y en el denominador probabilidad de que el suceso B ocurra. Esto muestra que la probabilidad condicional es una fracción de la probabilidad del suceso que ya se realizó, y esta fracción nos muestra que proporción del evento también pertenece al evento A.

La definición de probabilidad condicional aproxima a la definición de probabilidad independientes, ¿Qué sucede si el evento A que queremos analizar no tiene elementos comunes con el evento B?. Esto implicaría que la probabilidad de que se realice A dado que ya se realizó B es la misma que la probabilidad de se realice A, por lo tanto la probabilidad condicional es igual a la probabilidad del evento, como se evidencia a continuación:

$$P(A|B)=P(A)$$

Para demostrar este resultado, recordemos que pasa cuando deseamos calcular la probabilidad de que el evento A y B ocurran simultáneamente, pero ambos eventos son mutuamente excluyentes. Que ambos evento ocurran simultáneamente es lo mismo que exigir que elementos del evento A también estén en el suceso B, por lo tanto la formula sería:

$$P(A\ y\ B)=P(A\cap B)=P(A)\,x\,P(B)$$

Introduciendo esto en la fórmula de la probabilidad condicional tenemos que:

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

Dado que en el numerador esta una multiplicación, ocupando las reglas de la simplificación de fracciones, podemos despejar y encontrar el resultado esperado.

$$P(A|B)=P(A)$$

Esto muestra que la dependencia o independencia que exista entre los sucesos condicionará el cálculo de la probabilidad que deseamos. Debemos ser minuciosos al momento de saber si el evento que analizamos depende de un evento que se realizará con antelación o no para no incurrir en errores metodológicos al momento de calcular la probabilidad.

### 3.5.2- Teorema de la probabilidad total

El teorema de la probabilidad total nos muestra como calcular la probabilidad de un suceso, dado que conocemos sus probabilidades condicionales. Esto implica conocer la probabilidad condicional del evento B dado una serie de eventos que se han realizado ya. Lo que deseamos es conocer la probabilidad de que el evento B suceda, no su probabilidad condicionada.

Pensemos en el siguiente ejemplo, podemos conocer la probabilidad de que nos duela la cabeza dado que estamos resfriados, o que nos duela la cabeza dado que estamos estresados, o que nos duela la cabeza dado que tenemos insomnio. Conocemos tres probabilidades condicionadas para un mismo suceso, que nos duela la cabeza, y deseamos conocer su probabilidad de realización. El teorema de la probabilidad total nos permite esto.

Debemos puntualizar que los sucesos que condicionen el suceso que estamos analizando deben ser parte del mismo espacio muestral, en el ejemplo que dimos los tres sucesos son posibles enfermedades, por lo tanto pertenecen al espacio muestra que calcula la probabilidad de tener alguna enfermedad. Formalmente el teorema es el siguiente:

***Teorema de la Probabilidad total**: Sea* $A_1, A_2, \ldots\ldots, A_n$ *elementos que pertenecen al espacio muestral de un suceso y sea B un suceso cualquiera del que se conocen solo sus probabilidades condicionales* $P(B \vee A_i)$ *, entonces la probabilidad del suceso B será:*

$$P(B) = \sum_{i=1}^{n} P(B \vee A_i) P(A_i)$$

Esto nos muestra que la probabilidad del suceso será la suma la multiplicación de cada probabilidad condicional por la probabilidad del suceso que está condicionando, o sea, para el ejemplo dado:

$$P(B) = P(B|resfrio)\, x\, P(resfrio) + P(B|estres)\, x\, P(estres) + P(B|insomnio)\, x\, P(insomnio)$$

Esto nos ilustra como poder obtener la probabilidad de sucesos que están condicionados, o que son difíciles de experimentar de forma independiente.

## 3.6- Teorema de bayes

El teorema de bayes busca generar una probabilidad "*actualizada*" dado los contextos, esto quiere decir que intentamos generar una probabilidad para un suceso que dependió de la realización de otro. En particular podemos pensar en que un suceso puede ser A1 o A2, luego de verificar cuál de los sucesos se realizó se desencadena el suceso B que es común a ambos.

Pensemos en que el suceso B es un dolor de cabeza, A1 es un resfrió y A2 una gripe, yo deseo calcular la probabilidad condicionada de que tenga gripe dado que me duele la

cabeza, pero existe una probabilidad de tener gripe y de estar resfriado. Vemos que A1 y A2 son mutuamente excluyentes y colectivamente exhaustivos, o sea que la realización de uno implica que el otro no se realiza y que siempre uno de los dos se dará.

Para saber la probabilidad que buscamos es necesario intentar hacer una "*proporción*" de probabilidades, o sea establecer que posibilidades tengo de que este enfermo, luego de que esa enfermedad sea gripe y que me duela la cabeza por tener gripe, mediante este teorema podremos aproximarnos a esto. La fórmula del **teorema de bayes** es el siguiente:

$$P(A_1|B)=\frac{P(A_1)P(B\vee A_1)}{P(A_1)P(B|A_1)+P(A_2)P(B\vee A_2)}$$

La probabilidad que se calcula sin tener una certeza adicional del experimento, o sea antes de la realización del mismo, se denomina ***probabilidad a priori*** e intenta obtener información a partir de los datos que se tienen actualmente (antes de la realización del suceso). Esta probabilidad seria la probabilidad de estar resfriado o con gripe, en el marco de la fórmula sería $P(A_1)$ o $P(A_2)$ .

Luego de ocupar información adicional, o sea sentir el dolor de cabeza y entender que estoy enfermo, puedo calcular una probabilidad "*actualizada*" con la información adicional que obtuve. Esta probabilidad "*actualizada*" recibe el nombre de ***probabilidad a posteriori*** y sería la probabilidad de tener gripe dado que me duele la cabeza, en la formula sería $P(A_1|B)$ .

Si analizamos la fórmula, veremos que en el numerador de la fracción esta la probabilidad de que me duela la cabeza y tenga gripe (recuerde la *Regla general de la multiplicación*) mientras tanto en el denominado esta la probabilidad de tener dolor de cabeza dado que estoy resfriado sumado a la probabilidad de tener dolor de cabeza dado que tengo gripe. Vemos que en el denominador tenemos el conjunto total de probabilidades estamos ocupando la *Regla general de la adición* para calcular la probabilidad de que me duela la cabeza debido a que tengo gripe o a que estoy resfriado.

Esto nos muestra que el *teorema de bayes* se basa en dividir la probabilidad de llegar a un punto por el conjunto anterior a él, en este caso la probabilidad de tener dolor de cabeza dado que tengo gripe dividido la probabilidad de tener dolor de cabeza dado que estoy enfermo.

Esto es sumamente útil debido a que obtener la probabilidad a posteriori es bastante difícil de forma experimental, debido a que el suceso por el cual condicionamos es el segundo que pasa, generalmente, por lo tanto es complejo imaginar un experimento diseñado así.

# CAPÍTULO CUATRO

# Distribuciones

El siguiente capítulo busca mostrar la forma gráfica en que se pueden distribuir distintas variables, en específico las que se desprenden de un ejercicio de probabilidades. Para las Ciencias Sociales es fundamental conocer tanto las distribuciones, pues permite dar el paso de la simple descripción al análisis y conclusión de diferentes experimentos que pueden realizarse o trabajarse.

Concluido el capítulo se espera que usted logre los siguientes objetivos:

- Comprender el significado de una variable aleatoria
- Entender y calcular la Media, Varianza y Desviación estándar de una variable
- Diferenciar Distribuciones Discretas de las Continuas
- Conocer distintas formas de Distribuciones Discretas
- Conocer distintas formas de Distribuciones Continuas

## 4.1 – Introducción

La realización de un experimento cualquiera nos arrojará un resultado en particular, pero si repetimos el experimento un par de veces más podríamos obtener resultados distintos cada vez. Como vimos anteriormente las probabilidades nos ayudan a acercarnos al resultado específico que deseamos obtener, pero la pregunta es cómo se distribuyen la totalidad de resultados que hemos obtenido al repetir muchas veces un experimento.

Una distribución es la forma gráfica en que se van obteniendo los resultados de un experimento en particular al repetirlo muchas veces, la gracia de esto es que nos permite ver las probabilidades asociadas a los posibles resultados que podemos obtener. Por lo tanto podemos definir una distribución de probabilidades

***Distribución de Probabilidad****: Listado de resultados de un experimento, al reiterarlo n veces, con la probabilidad de suceso al que está asociado.*

Esto nos muestra que una distribución nos mostrará como se conforma un espacio muestral, mostrando cada uno de sus elementos asociado a la probabilidad de suceso. Lo anterior terminará caracterizando al espacio muestral.

Una distribución de probabilidad de be cumplir las siguientes características para ser denominada como tal:

- La probabilidad de un resultado en particular se encuentra entre 0 y 1, inclusive
- Los resultados son mutuamente excluyentes
- La lista es exhaustiva, esto implica que la suma de las probabilidades de los distintos eventos es 1.

A lo largo del capítulo desarrollaremos las cualidades y los tipos de distribución que existen, pero partiremos con dos apartados básicos para entenderlas, primero hablaremos de las variables aleatorias y luego de la media, varianza y desviación estándar. A continuación veremos los dos grandes tipos de distribuciones, las discretas y las continuas con ejemplos en cada una de ellas.

## 4.2 – Variables Aleatorias

La variable aleatoria proviene del resultado de un experimento aleatorio, recordemos que estos últimos se definen así porque sus resultados dependen del azar, o no existe una certeza a priori del resultado final. Esto nos dice que la variable aleatoria dependerá del azar y por lo tanto podemos asignar alguna probabilidad a su ocurrencia.

***Variable Aleatoria***: *Resultado de un experimento aleatorio. Son resultados que se dan por la intervención del azar, por lo tanto, pueden adoptar distintos valores*

Podemos encontrar dos tipos de variables aleatoria, los cuales dependerán de como se conformen sus elementos. Primero tenemos las **variables aleatorias discretas**, las cuales tienen elementos claramente diferenciables (o separables) y **las variables aleatorias continúas** las cuales se componen por elementos continuos no separables.

La diferencia entre ambas, por lo tanto, será la separabilidad de las variables que las compongan. Si hablamos de estatura, peso o medidas de diámetro estamos frente a variables continuas, pero si hablamos de calificaciones o números enteros frente a variables discretas.

## 4.2 – Media, Varianza y Desviación Estándar

La media, varianza y desviación estándar son elementos fundamentales para poder caracterizar a cada distribución, y por lo tanto al espacio muestral que terminará graficando esta. Estas medidas nos ayudan a tener una visión numérica del ordenamiento de las variables en relación a su punto central, la media, y por lo tanto entender cuán lejos está algún valor de esta.

Puntualmente, diremos que la **media** es el valor central de la distribución de probabilidad, por lo tanto será el valor que se encuentra "*en el medio*" de la distribución. En el caso de las distribuciones de probabilidad, la media será también el valor promedio de larga duración de una variable aleatoria y el ***valor esperado*** de la misma. Este último elemento es un símil al promedio en variables numéricas, pero en el caso de las probabilidades debemos multiplicar el resultado por la probabilidad de ocurrencia, el resultado de esto será el valor esperado de la variable. Concretamente la fórmula de la media será la siguiente:

$$Media: \mu = \sum \left[ xP(x) \right]$$

Esto nos muestra que para calcular la media debemos multiplicar cada observación por la probabilidad de su ocurrencia y sumar estos resultados.

La **varianza** hace referencia al grado de dispersión que tienen las observaciones de una variable aleatoria. Entenderemos por dispersión la diferencia que existe entre la media y la observación puntual, esto nos mostrara cuan "*lejos*" se encuentran las observaciones de la media. La fórmula de la varianza es la siguiente:

$$Varianza: \sigma^2 = \sum \left[ (x-\mu)^2 P(x) \right]$$

Vemos que la diferencia entre la observación y la media se eleva al cuadrado, esto tiene dos explicaciones: primero, al elevar al cuadrado producimos que la diferencia sea siempre positiva otorgando una noción espacial a esta diferencia; segundo, cuando elevamos al cuadrado "*penalizamos*" las diferencias más grandes.

Las propiedades antes descritas terminan generando que la varianza sea un indicador de dispersión, lo que equivale a decir que nos aporta una noción de distancia o separación. Esto es importante al momento de analizar la distribución, pues, no es lo mismo aglutinar todas las observaciones que tenerlas muy dispersas dentro de un plano dado.

Finalmente, la **desviación estándar** proviene del cálculo de la varianza, y se dirá que es la distancia promedio que existe entre una observación y la media. En términos de la obtención numérica de ésta, diremos que es la raíz cuadrada de la varianza. La fórmula de la desviación estándar se presenta a continuación:

$$Desviación\,Estándar: \sigma = \sqrt{\sigma^2}$$

Con estos tres elementos ya podemos caracterizar cualquier distribución y obtener información de ella sin, necesariamente, conocer la fórmula que la genera.

## 4.3 – Distribuciones Discretas

Como ya se enunció anteriormente, una distribución de probabilidades discreta hace referencia a la localización dentro de un plano de una variable aleatoria discreta. Lo anterior implica que tenemos una variable aleatoria, o sea que sus resultados no tienen certeza absoluta, que es de orden discreta, o que sus elementos son separables.

La separabilidad de sus elementos hace referencia a que no podemos graficarlos de forma continua, sino como puntos dentro de un plano cartesiano. Esto implica que la variable responde a valores numéricos separables, como lo son los números enteros.

Que un valor sea separable hace referencia a la existencia de un espacio, podríamos decir que es un vacío, entre dos elementos de un grupo. Si pensamos en los números naturales vemos que el 5 y 6 pertenecen a este conjunto, ¿Qué existe entre el número 5 y el número 6 en el conjunto de los números naturales?. Sabemos que los números naturales son un conjunto de enteros positivos que van desde cero hasta infinito, esto implica que entre dos números sucesivos que pertenezcan al conjunto de números naturales no exista un valor entero, vale decir entre 5 y 6, desde la óptima de los números naturales, existe un espacio.

A continuación procederemos a analizar tres tipos de Distribuciones Discretas, la distribución Binomial y la Hipergeométroca.

### 4.3.1 – Distribución Binomial

La distribución de probabilidad Binomial nace de una variable aleatoria que tiene dos posibles resultados, supongamos que es acierto o rechazo. Esto implica que tendremos dos posibles resultados asociados al experimento en cuestión, por lo tanto nuestro espacio muestral contendrá dos posibles elementos. Los requisitos para calificar a un experimento como binomial, y por consiguiente a su distribución como binomial, son los siguientes:

- El resultado de cada experimento se debe poder encasillar en dos posibles resultados que son mutuamente excluyentes.
- La variable aleatoria permite contar en una cantidad numerables, y discreta, los éxitos de cada posible resultado.
- La probabilidad de cada posible resultado es la misma en cada experimento.
- Los experimentos son independientes, esto implica que su realización no afecta el resultado de una realización futura.

Cada experimento por separado se conoce como "*Experimento de Bernoulli*" y hace referencia a un experimento con dos resultados posibles. Cuando la distribución binomial tiene un $n = 1$ diremos que la distribución es Bernoulli.

Para calcular una probabilidad binomial necesitamos corroborar lo siguiente:

1) El número de experimentos
2) La probabilidad de éxito de cada experimento

Lo anterior implica que debemos conocer cuántas veces replicaremos el experimento, y cuál es la probabilidad de que se dé un resultado en particular al realizar un experimento.

Supongamos el hecho de acertar un tiro al blanco, digamos que nos dan 10 opciones para tirar, cada tiro es independiente del anterior pues el tirar no influye la probabilidad de éxito del siguiente tiro. Para cada intento tengo las mismas opciones de resultado: Acertar y Fallar, cada una con una probabilidad general de ½ o de 0.5.

La fórmula para calcular una probabilidad binomial es la siguiente:

$$P(x)=\binom{n}{x}\pi^{x}(1-\pi)^{n-x}$$

Donde:

- $n$ es el número de pruebas
- $x$ es la variable aleatoria definida como éxitos
- $\pi$ es la probabilidad de éxito de cada resultado

La media y varianza para una binomial se obtienen con la siguiente fórmula

$$Media: \mu = n\pi$$

$$Varianza: \sigma^2 = n\pi(1-\pi)$$

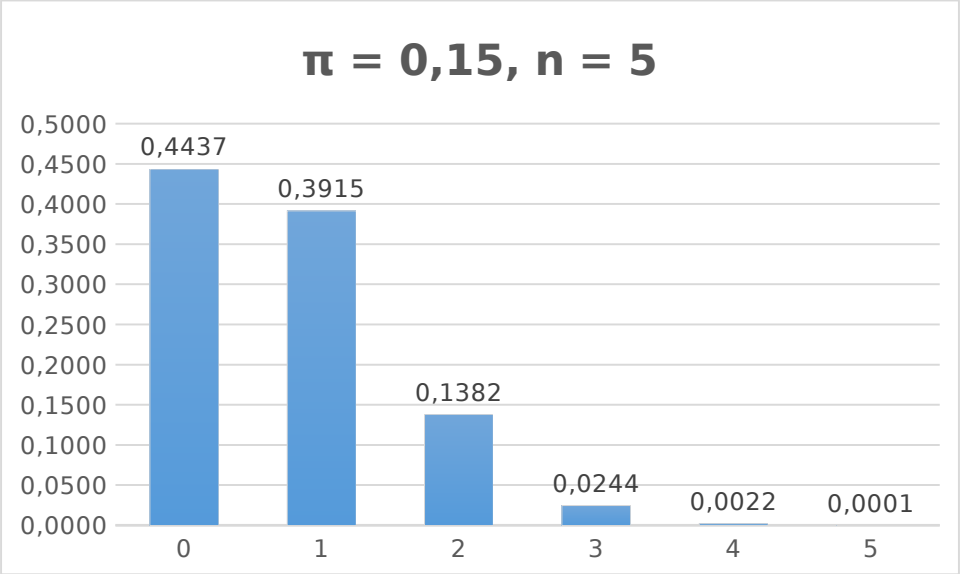
Mediante el cálculo de las probabilidades binomiales podemos obtener una tabla de probabilidades, y así obtener una distribución. A continuación se muestra una tabla de probabilidades binomial para un $n$ = 5 y distintas probabilidades de resultados.

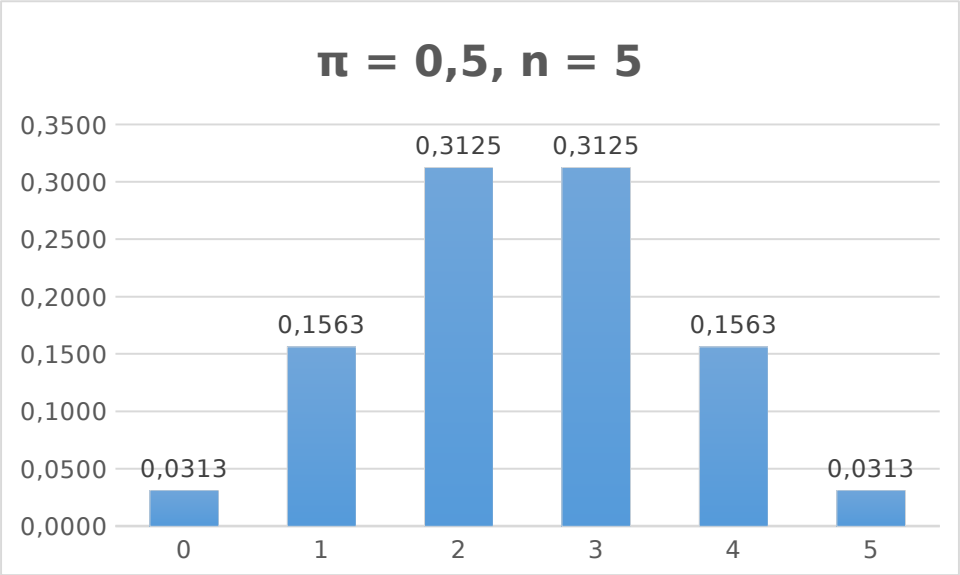
Probabilidad n = 5

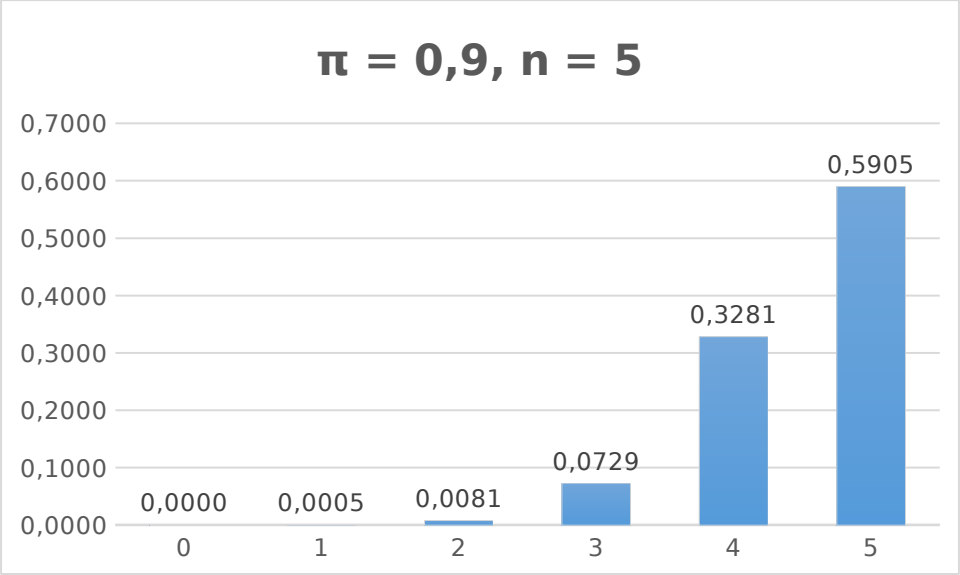
| x/π | 0,01 | 0,05 | 0,15 | 0,3 | 0,5 | 0,75 | 0,9 | 0,99 |
|---|---|---|---|---|---|---|---|---|
| 0 | 0,9510 | 0,7738 | 0,4437 | 0,1681 | 0,0313 | 0,0010 | 0,0000 | 0,0000 |
| 1 | 0,0480 | 0,2036 | 0,3915 | 0,3602 | 0,1563 | 0,0146 | 0,0005 | 0,0000 |
| 2 | 0,0010 | 0,0214 | 0,1382 | 0,3087 | 0,3125 | 0,0879 | 0,0081 | 0,0000 |
| 3 | 0,0000 | 0,0011 | 0,0244 | 0,1323 | 0,3125 | 0,2637 | 0,0729 | 0,0010 |
| 4 | 0,0000 | 0,0000 | 0,0022 | 0,0284 | 0,1563 | 0,3955 | 0,3281 | 0,0480 |
| 5 | 0,0000 | 0,0000 | 0,0001 | 0,0024 | 0,0313 | 0,2373 | 0,5905 | 0,9510 |

Notamos que la primera columna de la tabla tiene los mismos valores que la última, pero a la inversa.

A continuación se mostrara la distribución gráfica de alguna de las probabilidades que están en la tabla.



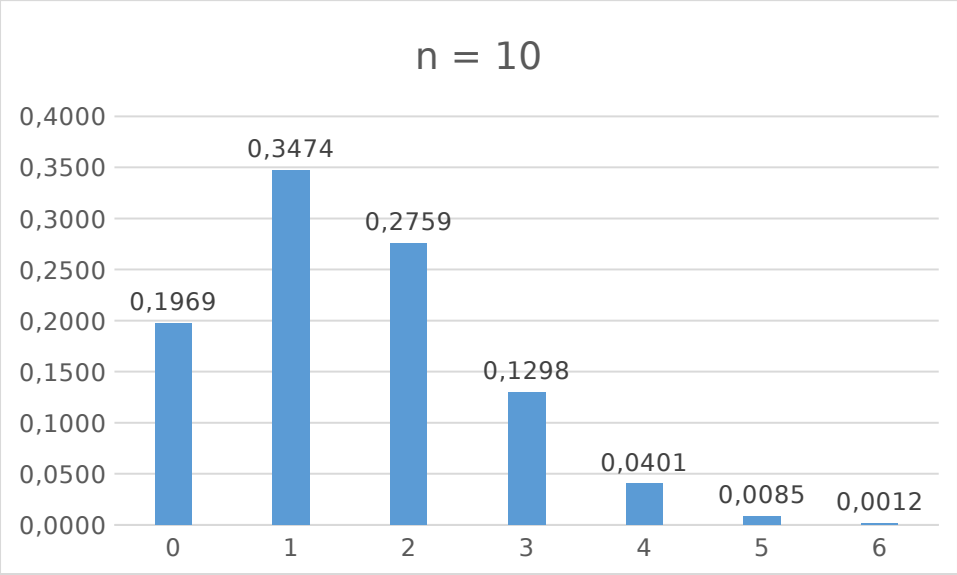


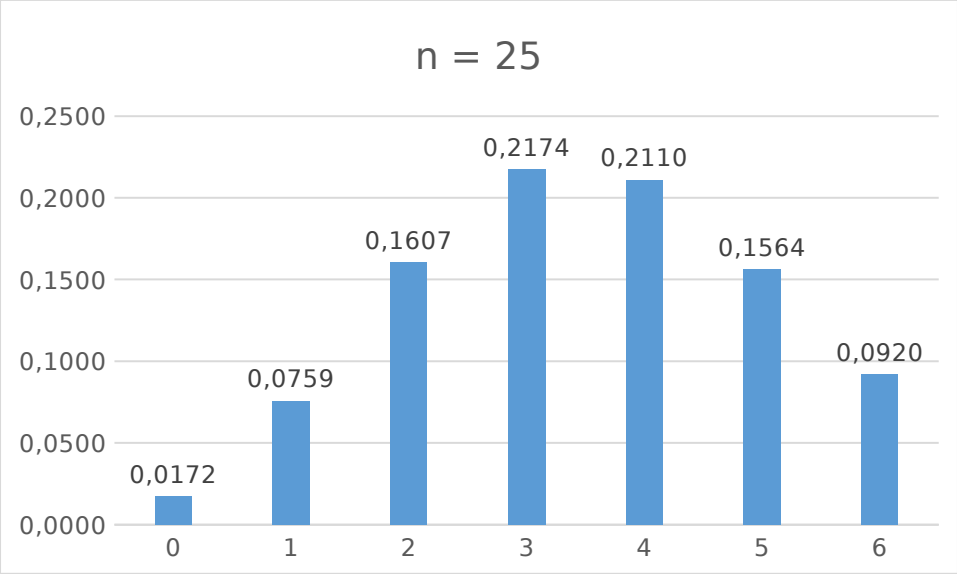

Lo anterior es como se mueve una distribución cuando mantenemos la cantidad de experimentos pero alteramos la probabilidad de éxito y el número de aciertos. Comprobamos que la distribución va cambiando su forma, puntualmente pareciera que se traslada desde un extremo al otro. Veamos que sucede si mantenemos el número de éxitos y la probabilidad de éxito pero incrementamos el número de experimentos.

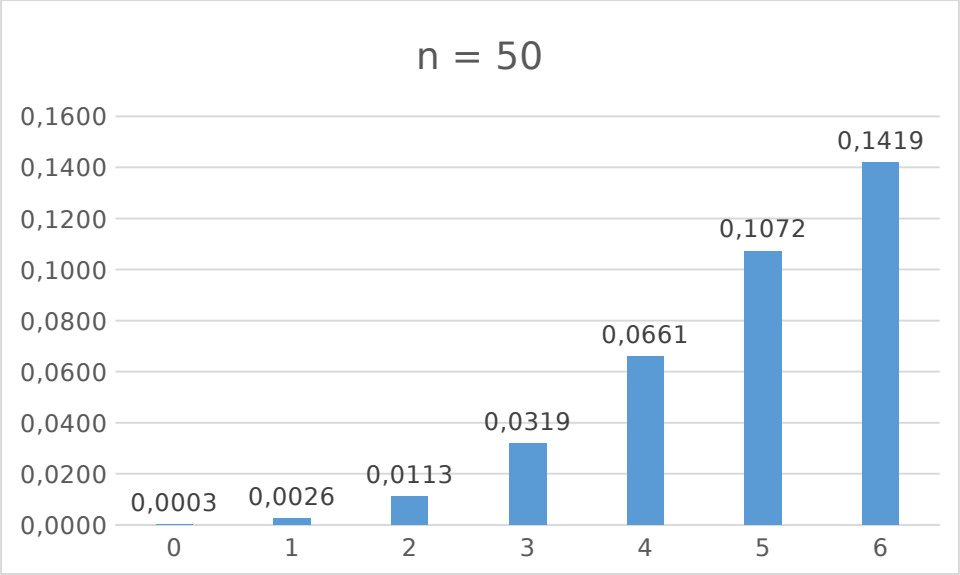
Probabilidad π = 0,15

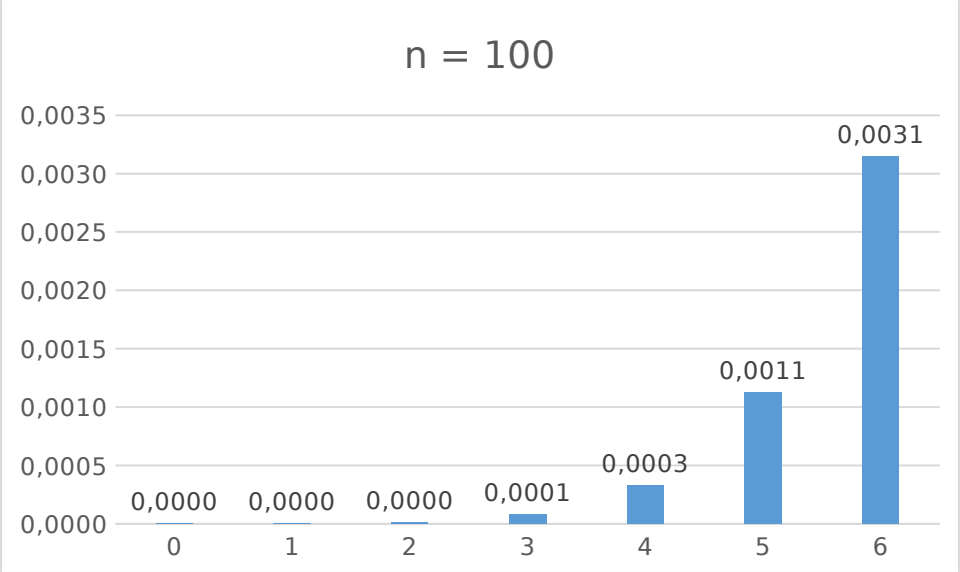
| x/n | 10 | 25 | 50 | 100 |
|---|---|---|---|---|
| 0 | 0,1969 | 0,0172 | 0,0003 | 0,0000 |
| 1 | 0,3474 | 0,0759 | 0,0026 | 0,0000 |
| 2 | 0,2759 | 0,1607 | 0,0113 | 0,0000 |
| 3 | 0,1298 | 0,2174 | 0,0319 | 0,0001 |
| 4 | 0,0401 | 0,2110 | 0,0661 | 0,0003 |
| 5 | 0,0085 | 0,1564 | 0,1072 | 0,0011 |
| 6 | 0,0012 | 0,0920 | 0,1419 | 0,0031 |

En este caso tenemos las probabilidades binomiales para un experimento que se repite 10, 25, 50 y 100 veces y que tiene una probabilidad de acierto de un 0.15 o 15%. Veamos las distribuciones de las estas probabilidades gráficamente.


n = 10
0,4000
0,3500
0,3000
0,2500
0,2000
0,1500
0,1000
0,0500
0,0000
0,1969
0,3474
0,2759
0,1298
0,0401
0,0085
0,0012
0
1
2
3
4
5
6


n = 25
0,2500
0,2000
0,1500
0,1000
0,0500
0,0000
0,0172
0,0759
0,1607
0,2174
0,2110
0,1564
0,0920
0
1
2
3
4
5
6




Vemos que a medida que aumenta el número de veces que repetimos el experimento la probabilidad asociada a cada realización de éxito disminuye y, a la ves, varía la forma de la distribución de probabilidad.

### 4.3.2 – Distribución Hipergeométrica

La distribución binomial se basa en que la probabilidad de acierto se mantiene constante en cada experimento, si tenemos 100 personas, 60 mujeres y 50 hombres, la probabilidad de "*sacar*" a una mujer es de un 60%. El tema es que si no existe reemplazo, o sea "devolver" a la persona a la muestra, en el segundo experimento ya no tendremos 100

personas sino 99 y de ellas 50 serán hombres y 59 mujeres, esto equivale a que la probabilidad de "*sacar*" una mujer ahora será 59,59%.

Frente a problemas así, en donde el experimento no permite reemplazos, no podemos ocupar la distribución binomial, es por esto que debemos buscar otro mecanismo para obtener la distribución de la variable. El mecanismo que ocuparemos para estos casos será la **Distribución Hipergeométrica**, esto implica que se cumplirán las condiciones para la distribución binomial, solo cambiará que la probabilidad de acierto varia a medida que aumentan los experimentos.

La fórmula de la probabilidad hipergeométrica es:

$$P(x) = \frac{\binom{S}{x}\binom{N-S}{n-x}}{\binom{N}{n}}$$

Donde:

- *N* es el tamaño de la población
- *S* es número de éxitos en la población
- *x* es el número de éxitos en la muestra
- *n* es el tamaño de la muestra o el número de pruebas

Una distribución de probabilidades hipergeométrica tiene las siguientes características:

- Los resultados de cada experimento se pueden calificar en dos categorías exclusivas
- La variable aleatoria es el número de éticos de un número fijo de experimentos
- Los experimentos no son independientes
- Los muestreos se realizan sin reemplazo y con una población finita con n/N < 0.05 lo que implica que las probabilidades cambian con cada experimento.

Veamos que sucede con la probabilidad hipergeométrica si fijamos el tamaño de la población en 100 y obtenemos una muestra de 20 personas.

Probabilidad N = 100, n = 20

| x/S | 30 | 50 | 80 |
|---|---|---|---|
| 1 | 0,004 | 0,000 | 0,000 |
| 5 | 0,192 | 0,009 | 0,000 |
| 10 | 0,022 | 0,197 | 0,001 |
| 15 | 0,000 | 0,009 | 0,192 |
| 20 | 0,000 | 0,000 | 0,007 |

Apreciamos que la primera columna hace referencia a la cantidad de éxitos en la muestra y la primera fila a la cantidad de éxitos en la población. Dado que esta distribución emerge

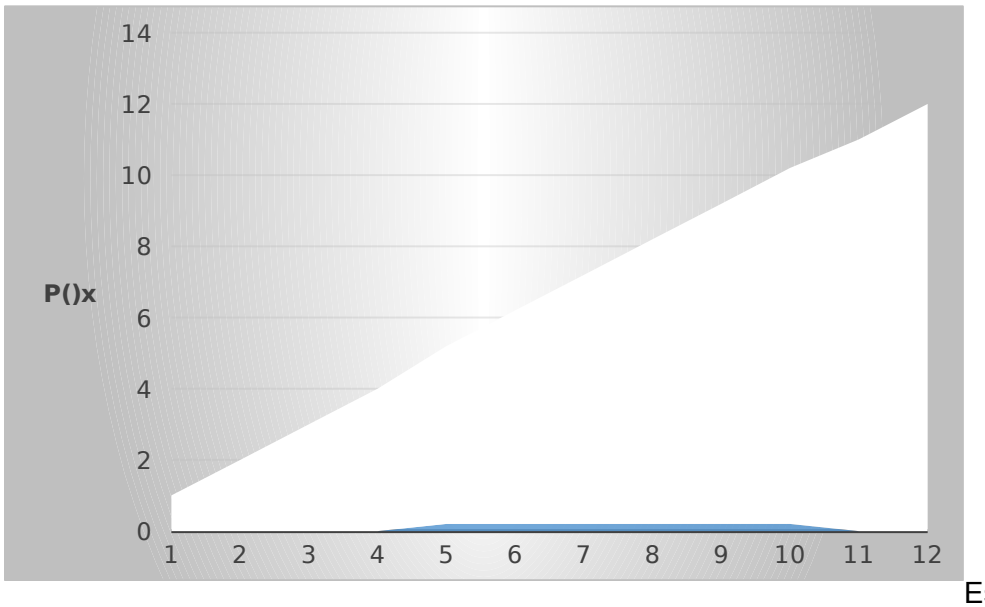
de la binomial, su conformación gráfica es idéntica a la que se describió en el apartado anterior.

## 4.4 – Distribuciones Continuas

La distribución de probabilidad continua se basa en variables aleatorias que son conformadas por elementos continuos. Un elemento continuo hace referencia a un evento que posee una métrica no divisible o separable, antes decíamos que los números naturales era un grupo separable, si decimos los números positivos veremos que "*agregamos*" a los naturales todo número decimal, generando que existan infinitos números entre el 5 y el 6 (esto es siguiendo nuestro ejemplo). Otro ejemplo sería pensar en la estatura de una persona, esta variable puede ser 1.7 metros, pero también 1.71 metros o 1.7111 metros, acá apreciamos la inseparabilidad de la variable.

A continuación analizaremos dos "*familias*" de distribuciones, la primera es la *distribución uniforme* que se caracteriza porque todos los elementos dentro del espacio muestral tiene la misma probabilidad de ocurrencia. La segunda familia hace referencia a la *distribución normal*, está proviene de la conocida campana de Gauss y es la más común de todas. Dentro de esta última familia, encontramos una estandarización clásica, la normal (0;1).

### 4.4.1 – Distribución Uniforme




Esta distribución tiene forma rectangular establecida entre un mínimo, denominado *a*, y un máximo, denominado *b*. La probabilidad de todos los elementos entre *a* y *b* será 1 partido en la restra entre ambas $\frac{1}{b-a}$ . A continuación mostramos un ejemplo gráfico para un mínimo de 5 y un máximo de 10 con la correspondiente probabilidad de 0,2.

A continuación se mostrará los valores para la media, desviación estándar y la probabilidad asociada a la distribución:

$$Media: \mu = \frac{a+b}{2}$$

$$Desviación\ estándar: \sigma = \sqrt{\frac{(b-a)^2}{12}}$$

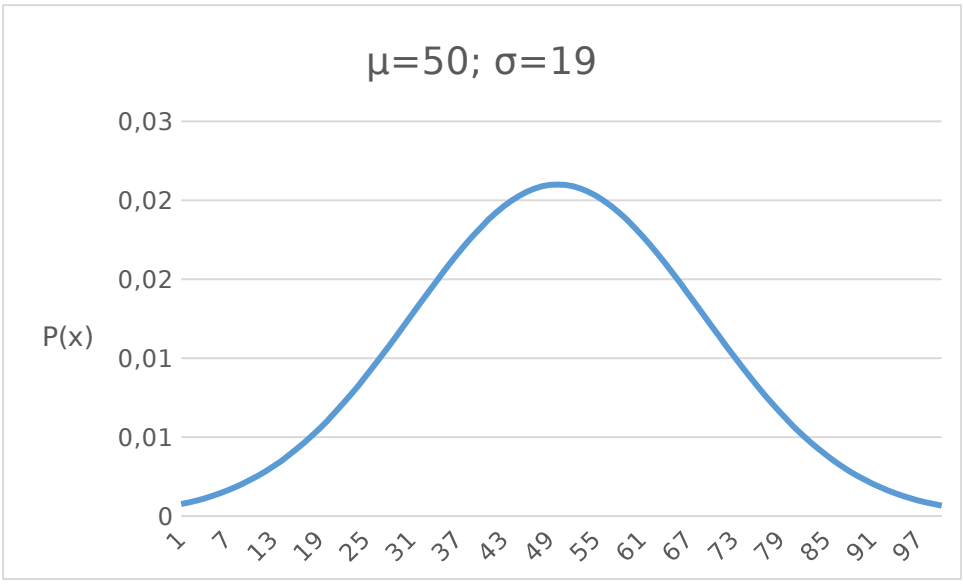
$$Probabilidad\ de\ la\ distribución\ Uniforme: P(x) = \frac{1}{b-a}\ ;si\ a \leq x \leq b$$

En cualquier punto que no pertenezca al intervalo entre *a* y *b* la probabilidad será 0. Apreciamos que esta distribución es bastante sencilla, debido a su característica de uniforme, pero debemos tener cuidado en saber cuando aplicarla.

Por ejemplo, pensemos en la probabilidad de abordar una movilización pública, para el caso de Chile un *Transantiago* o sus homologables regionales, estos debiesen tener una frecuencia temporal dada, supongamos de 30 minutos. Si los pasajeros llegan a la parada aleatoriamente, todos los que lleguen entre los 0 y 30 minutos tienen la misma probabilidad de abordar el bus, esto debido a que aún no pasa por lo tanto lo deben esperar. Este ejemplo nos demuestra un caso bajo el cual podemos aplicar la distribución que acabamos de ver.

### 4.4.2 – Distribución Normal

La distribución normal se define en base a su media y varianza, esto quiere decir que la distribución gráfica viene dada por estos dos elementos. Esta distribución es la que más se aprecia en la observación de hechos en la vida real, esto debido a que permite que la "*mayoría*" de los hechos se concentren alrededor de la media, teniendo pocas variables que se separen de esta. Gráficamente una distribución normal es de la siguiente manera:




Este gráfico muestra la distribución normal para una variable de N=100, una media de 50 y una desviación estándar de 19. Apreciamos que el valor máximo de la distribución se da en el la media de la misma y que la mayoría de los datos se agrupan en torno a la media, separándose en el valor de la desviación estándar, esto quiere decir que la mayoría de las observaciones se encuentran entre 31 y 69. El eje vertical muestra la probabilidad de éxito de los sucesos.

Las características principales de esta distribución son:

- Tiene forma de campana y posee un máximo en el punto de la media aritmética. La media, moda y mediana son iguales en este caso y se localizan al centro de la distribución.
- Es simétrica respecto a su media, o sea que si separamos la distribución en su media, tendremos que los lados resultantes son idénticos.
- La distribución es asintótica al eje de las X, esto quiere decir que siempre existirán probabilidades de éxito positiva para valores que existan en el espacio muestral.
- La localización de la distribución se determina según su media, y su dispersión viene dada por la desviación estándar de esta.

La fórmula de las probabilidades que están detrás de esta distribución es la siguiente:

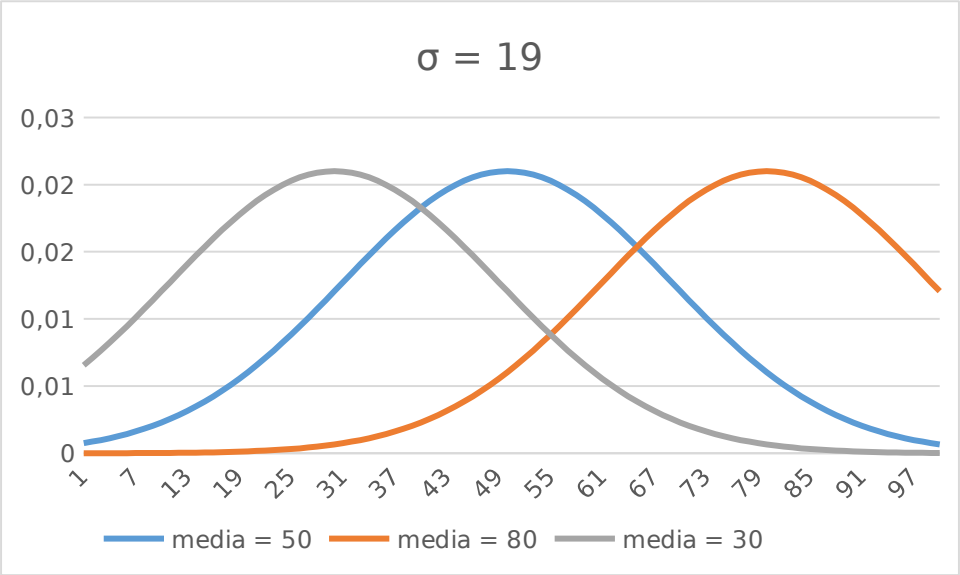
$$P(x): \frac{1}{\sigma\sqrt{2\pi}} e^{-\left[\frac{(X-\mu)^2}{2\sigma^2}\right]}$$

Donde:

- μ es la media de la distribución
- σ es la desviación estándar de la distribución
- π es la constante matemática igual a 3.1416

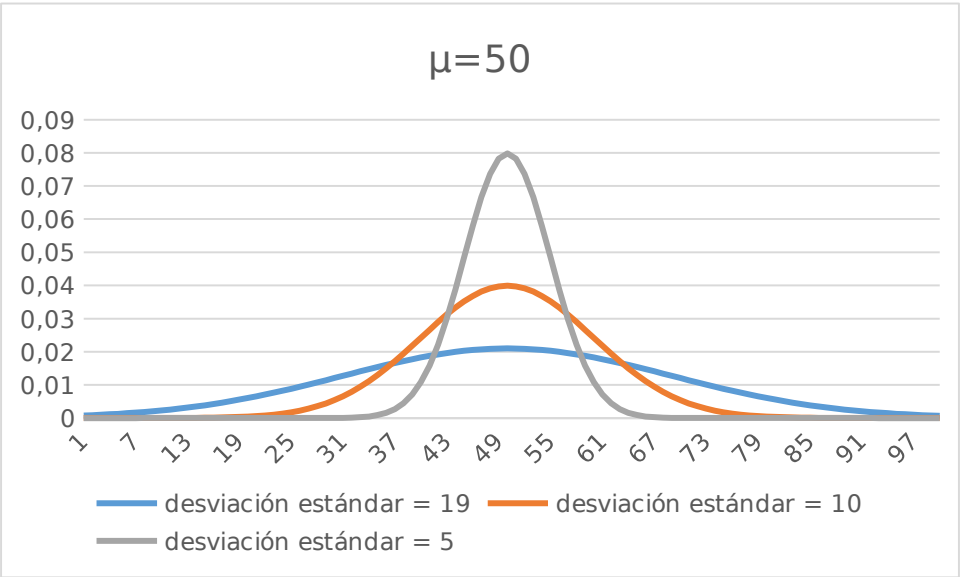
- e es la constante matemática exponencial igual a 2.718
- x es la observación que pertenece a la variable aleatoria.

Debido a que la distribución dependerá de su media y su desviación estándar, las variaciones en estas terminaran determinando como se verá la distribución. Frente a cambios en la media variará el punto máximo de la distribución, esto se aprecia en la siguiente figura.




Esta figura nos muestra tres distribuciones normales con distintas medias, pero igual desviación estándar. Vemos que se mantiene el punto máximo igual en las tres distribuciones, pero varia el punto en el cual se logra esta probabilidad máxima.

Apreciamos que sucede cuando hacemos variar la desviación estándar.



Vemos que la media sigue siendo la misma, 50, pero a medida que la desviación estándar baja la distribución tiende a tener una agrupación mayor en torno a la media, generando una probabilidad máxima más alta.

Un hecho singular que sucede con esta distribución es su normalización, esta hace referencia a una distribución normal con media 0 y desviación estándar 1, este tipo de distribución es conocido como ***distribución normal estándar***.

El término normalización hace referencia a "*transformar*" una distribución no normal, en una normal con las características antes dichas. Cualquier distribución normal puede llevarse hacia una normal estándar, y otras distribuciones que emergen desde la normal también. La fórmula para la normalización es la siguiente:

$$Valor\ normal\ estándar: Z = \frac{X - \mu}{\sigma}$$

Donde Z será el valor que conformará nuestra nueva distribución, vale decir, el valor que va en el eje de las Y o el equivalente a un nuevo espacio muestral.

Vemos que la normalización hace referencia a calcular la diferencia que tiene una observación en particular con la media de la distribución y dividir esto por la desviación estándar, es una medida de cuán lejos nos encontramos de la media y forzamos a que la desviación estándar sea 1.

Las probabilidades asociadas a estos valores, al igual que las asociadas a los valores para la media y desviación estándar de una distribución normal cualquiera, deben ser calculadas según la formula descrita arriba. Generalmente estos valores se muestran en tablas de distribución.

A continuación se presentan dos tablas que muestran las probabilidades de una normal estándar, recordar que la normal es una distribución simétrica, por lo tanto, como lo verá en las tablas, las probabilidades de un número positivo son las mismas que de los negativos.

Tabla Distribución Normal Uniforme, números positivos

| z | 0,01 | 0,02 | 0,03 | 0,04 | 0,05 | 0,06 | 0,07 | 0,08 | 0,09 |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 0,3989 | 0,3989 | 0,3988 | 0,3986 | 0,3984 | 0,3982 | 0,3980 | 0,3977 | 0,3973 |
| 0,1 | 0,3965 | 0,3961 | 0,3956 | 0,3951 | 0,3945 | 0,3939 | 0,3932 | 0,3925 | 0,3918 |
| 0,2 | 0,3902 | 0,3894 | 0,3885 | 0,3876 | 0,3867 | 0,3857 | 0,3847 | 0,3836 | 0,3825 |
| 0,3 | 0,3802 | 0,3790 | 0,3778 | 0,3765 | 0,3752 | 0,3739 | 0,3725 | 0,3712 | 0,3697 |
| 0,4 | 0,3668 | 0,3653 | 0,3637 | 0,3621 | 0,3605 | 0,3589 | 0,3572 | 0,3555 | 0,3538 |
| 0,5 | 0,3503 | 0,3485 | 0,3467 | 0,3448 | 0,3429 | 0,3410 | 0,3391 | 0,3372 | 0,3352 |
| 0,6 | 0,3312 | 0,3292 | 0,3271 | 0,3251 | 0,3230 | 0,3209 | 0,3187 | 0,3166 | 0,3144 |
| 0,7 | 0,3101 | 0,3079 | 0,3056 | 0,3034 | 0,3011 | 0,2989 | 0,2966 | 0,2943 | 0,2920 |
| 0,8 | 0,2874 | 0,2850 | 0,2827 | 0,2803 | 0,2780 | 0,2756 | 0,2732 | 0,2709 | 0,2685 |
| 0,9 | 0,2637 | 0,2613 | 0,2589 | 0,2565 | 0,2541 | 0,2516 | 0,2492 | 0,2468 | 0,2444 |
| 1 | 0,2396 | 0,2371 | 0,2347 | 0,2323 | 0,2299 | 0,2275 | 0,2251 | 0,2227 | 0,2203 |
| 1,1 | 0,2155 | 0,2131 | 0,2107 | 0,2083 | 0,2059 | 0,2036 | 0,2012 | 0,1989 | 0,1965 |
| 1,2 | 0,1919 | 0,1895 | 0,1872 | 0,1849 | 0,1826 | 0,1804 | 0,1781 | 0,1758 | 0,1736 |
| 1,3 | 0,1691 | 0,1669 | 0,1647 | 0,1626 | 0,1604 | 0,1582 | 0,1561 | 0,1539 | 0,1518 |
| 1,4 | 0,1476 | 0,1456 | 0,1435 | 0,1415 | 0,1394 | 0,1374 | 0,1354 | 0,1334 | 0,1315 |
| 1,5 | 0,1276 | 0,1257 | 0,1238 | 0,1219 | 0,1200 | 0,1182 | 0,1163 | 0,1145 | 0,1127 |
| 1,6 | 0,1092 | 0,1074 | 0,1057 | 0,1040 | 0,1023 | 0,1006 | 0,0989 | 0,0973 | 0,0957 |
| 1,7 | 0,0925 | 0,0909 | 0,0893 | 0,0878 | 0,0863 | 0,0848 | 0,0833 | 0,0818 | 0,0804 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1,8 | 0,0775 | 0,0761 | 0,0748 | 0,0734 | 0,0721 | 0,0707 | 0,0694 | 0,0681 | 0,0669 |
| 1,9 | 0,0644 | 0,0632 | 0,0620 | 0,0608 | 0,0596 | 0,0584 | 0,0573 | 0,0562 | 0,0551 |
| 2 | 0,0529 | 0,0519 | 0,0508 | 0,0498 | 0,0488 | 0,0478 | 0,0468 | 0,0459 | 0,0449 |
| 2,1 | 0,0431 | 0,0422 | 0,0413 | 0,0404 | 0,0396 | 0,0387 | 0,0379 | 0,0371 | 0,0363 |
| 2,2 | 0,0347 | 0,0339 | 0,0332 | 0,0325 | 0,0317 | 0,0310 | 0,0303 | 0,0297 | 0,0290 |
| 2,3 | 0,0277 | 0,0270 | 0,0264 | 0,0258 | 0,0252 | 0,0246 | 0,0241 | 0,0235 | 0,0229 |
| 2,4 | 0,0219 | 0,0213 | 0,0208 | 0,0203 | 0,0198 | 0,0194 | 0,0189 | 0,0184 | 0,0180 |
| 2,5 | 0,0171 | 0,0167 | 0,0163 | 0,0158 | 0,0154 | 0,0151 | 0,0147 | 0,0143 | 0,0139 |
| 2,6 | 0,0132 | 0,0129 | 0,0126 | 0,0122 | 0,0119 | 0,0116 | 0,0113 | 0,0110 | 0,0107 |
| 2,7 | 0,0101 | 0,0099 | 0,0096 | 0,0093 | 0,0091 | 0,0088 | 0,0086 | 0,0084 | 0,0081 |
| 2,8 | 0,0078 | 0,0075 | 0,0073 | 0,0071 | 0,0069 | 0,0067 | 0,0065 | 0,0063 | 0,0061 |
| 2,9 | 0,0058 | 0,0056 | 0,0055 | 0,0053 | 0,0051 | 0,0050 | 0,0048 | 0,0047 | 0,0046 |
| 3 | 0,0043 | 0,0042 | 0,0040 | 0,0039 | 0,0038 | 0,0037 | 0,0036 | 0,0035 | 0,0034 |
| 3,1 | 0,0032 | 0,0031 | 0,0030 | 0,0029 | 0,0028 | 0,0027 | 0,0026 | 0,0025 | 0,0025 |
| 3,2 | 0,0023 | 0,0022 | 0,0022 | 0,0021 | 0,0020 | 0,0020 | 0,0019 | 0,0018 | 0,0018 |
| 3,3 | 0,0017 | 0,0016 | 0,0016 | 0,0015 | 0,0015 | 0,0014 | 0,0014 | 0,0013 | 0,0013 |
| 3,4 | 0,0012 | 0,0012 | 0,0011 | 0,0011 | 0,0010 | 0,0010 | 0,0010 | 0,0009 | 0,0009 |
| 3,5 | 0,0008 | 0,0008 | 0,0008 | 0,0008 | 0,0007 | 0,0007 | 0,0007 | 0,0007 | 0,0006 |
| 3,6 | 0,0006 | 0,0006 | 0,0005 | 0,0005 | 0,0005 | 0,0005 | 0,0005 | 0,0005 | 0,0004 |
| 3,7 | 0,0004 | 0,0004 | 0,0004 | 0,0004 | 0,0004 | 0,0003 | 0,0003 | 0,0003 | 0,0003 |
| 3,8 | 0,0003 | 0,0003 | 0,0003 | 0,0003 | 0,0002 | 0,0002 | 0,0002 | 0,0002 | 0,0002 |
| 3,9 | 0,0002 | 0,0002 | 0,0002 | 0,0002 | 0,0002 | 0,0002 | 0,0002 | 0,0001 | 0,0001 |
| 4 | 0,0001 | 0,0001 | 0,0001 | 0,0001 | 0,0001 | 0,0001 | 0,0001 | 0,0001 | 0,0001 |

Tabla Distribución Normal Uniforme, números negativos

| z | 0,01 | 0,02 | 0,03 | 0,04 | 0,05 | 0,06 | 0,07 | 0,08 | 0,09 |
|---|---|---|---|---|---|---|---|---|---|
| -4 | 0,0001 | 0,0001 | 0,0002 | 0,0002 | 0,0002 | 0,0002 | 0,0002 | 0,0002 | 0,0002 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| -3,9 | 0,0002 | 0,0002 | 0,0002 | 0,0002 | 0,0002 | 0,0003 | 0,0003 | 0,0003 | 0,0003 |
| -3,8 | 0,0003 | 0,0003 | 0,0003 | 0,0003 | 0,0004 | 0,0004 | 0,0004 | 0,0004 | 0,0004 |
| -3,7 | 0,0004 | 0,0005 | 0,0005 | 0,0005 | 0,0005 | 0,0005 | 0,0005 | 0,0006 | 0,0006 |
| -3,6 | 0,0006 | 0,0007 | 0,0007 | 0,0007 | 0,0007 | 0,0008 | 0,0008 | 0,0008 | 0,0008 |
| -3,5 | 0,0009 | 0,0009 | 0,0010 | 0,0010 | 0,0010 | 0,0011 | 0,0011 | 0,0012 | 0,0012 |
| -3,4 | 0,0013 | 0,0013 | 0,0014 | 0,0014 | 0,0015 | 0,0015 | 0,0016 | 0,0016 | 0,0017 |
| -3,3 | 0,0018 | 0,0018 | 0,0019 | 0,0020 | 0,0020 | 0,0021 | 0,0022 | 0,0022 | 0,0023 |
| -3,2 | 0,0025 | 0,0025 | 0,0026 | 0,0027 | 0,0028 | 0,0029 | 0,0030 | 0,0031 | 0,0032 |
| -3,1 | 0,0034 | 0,0035 | 0,0036 | 0,0037 | 0,0038 | 0,0039 | 0,0040 | 0,0042 | 0,0043 |
| -3 | 0,0046 | 0,0047 | 0,0048 | 0,0050 | 0,0051 | 0,0053 | 0,0055 | 0,0056 | 0,0058 |
| -2,9 | 0,0061 | 0,0063 | 0,0065 | 0,0067 | 0,0069 | 0,0071 | 0,0073 | 0,0075 | 0,0077 |
| -2,8 | 0,0081 | 0,0084 | 0,0086 | 0,0088 | 0,0091 | 0,0093 | 0,0096 | 0,0099 | 0,0101 |
| -2,7 | 0,0107 | 0,0110 | 0,0113 | 0,0116 | 0,0119 | 0,0122 | 0,0126 | 0,0129 | 0,0132 |
| -2,6 | 0,0139 | 0,0143 | 0,0147 | 0,0151 | 0,0154 | 0,0158 | 0,0163 | 0,0167 | 0,0171 |
| -2,5 | 0,0180 | 0,0184 | 0,0189 | 0,0194 | 0,0198 | 0,0203 | 0,0208 | 0,0213 | 0,0219 |
| -2,4 | 0,0229 | 0,0235 | 0,0241 | 0,0246 | 0,0252 | 0,0258 | 0,0264 | 0,0270 | 0,0277 |
| -2,3 | 0,0290 | 0,0297 | 0,0303 | 0,0310 | 0,0317 | 0,0325 | 0,0332 | 0,0339 | 0,0347 |
| -2,2 | 0,0363 | 0,0371 | 0,0379 | 0,0387 | 0,0396 | 0,0404 | 0,0413 | 0,0422 | 0,0431 |
| -2,1 | 0,0449 | 0,0459 | 0,0468 | 0,0478 | 0,0488 | 0,0498 | 0,0508 | 0,0519 | 0,0529 |
| -2 | 0,0551 | 0,0562 | 0,0573 | 0,0584 | 0,0596 | 0,0608 | 0,0620 | 0,0632 | 0,0644 |
| -1,9 | 0,0669 | 0,0681 | 0,0694 | 0,0707 | 0,0721 | 0,0734 | 0,0748 | 0,0761 | 0,0775 |
| -1,8 | 0,0804 | 0,0818 | 0,0833 | 0,0848 | 0,0863 | 0,0878 | 0,0893 | 0,0909 | 0,0925 |
| -1,7 | 0,0957 | 0,0973 | 0,0989 | 0,1006 | 0,1023 | 0,1040 | 0,1057 | 0,1074 | 0,1092 |
| -1,6 | 0,1127 | 0,1145 | 0,1163 | 0,1182 | 0,1200 | 0,1219 | 0,1238 | 0,1257 | 0,1276 |
| -1,5 | 0,1315 | 0,1334 | 0,1354 | 0,1374 | 0,1394 | 0,1415 | 0,1435 | 0,1456 | 0,1476 |
| -1,4 | 0,1518 | 0,1539 | 0,1561 | 0,1582 | 0,1604 | 0,1626 | 0,1647 | 0,1669 | 0,1691 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| -1,3 | 0,1736 | 0,1758 | 0,1781 | 0,1804 | 0,1826 | 0,1849 | 0,1872 | 0,1895 | 0,1919 |
| -1,2 | 0,1965 | 0,1989 | 0,2012 | 0,2036 | 0,2059 | 0,2083 | 0,2107 | 0,2131 | 0,2155 |
| -1,1 | 0,2203 | 0,2227 | 0,2251 | 0,2275 | 0,2299 | 0,2323 | 0,2347 | 0,2371 | 0,2396 |
| -1 | 0,2444 | 0,2468 | 0,2492 | 0,2516 | 0,2541 | 0,2565 | 0,2589 | 0,2613 | 0,2637 |
| -0,9 | 0,2685 | 0,2709 | 0,2732 | 0,2756 | 0,2780 | 0,2803 | 0,2827 | 0,2850 | 0,2874 |
| -0,8 | 0,2920 | 0,2943 | 0,2966 | 0,2989 | 0,3011 | 0,3034 | 0,3056 | 0,3079 | 0,3101 |
| -0,7 | 0,3144 | 0,3166 | 0,3187 | 0,3209 | 0,3230 | 0,3251 | 0,3271 | 0,3292 | 0,3312 |
| -0,6 | 0,3352 | 0,3372 | 0,3391 | 0,3410 | 0,3429 | 0,3448 | 0,3467 | 0,3485 | 0,3503 |
| -0,5 | 0,3538 | 0,3555 | 0,3572 | 0,3589 | 0,3605 | 0,3621 | 0,3637 | 0,3653 | 0,3668 |
| -0,4 | 0,3697 | 0,3712 | 0,3725 | 0,3739 | 0,3752 | 0,3765 | 0,3778 | 0,3790 | 0,3802 |
| -0,3 | 0,3825 | 0,3836 | 0,3847 | 0,3857 | 0,3867 | 0,3876 | 0,3885 | 0,3894 | 0,3902 |
| -0,2 | 0,3918 | 0,3925 | 0,3932 | 0,3939 | 0,3945 | 0,3951 | 0,3956 | 0,3961 | 0,3965 |
| -0,1 | 0,3973 | 0,3977 | 0,3980 | 0,3982 | 0,3984 | 0,3986 | 0,3988 | 0,3989 | 0,3989 |
| 0 | 0,3989 | 0,3989 | 0,3988 | 0,3986 | 0,3984 | 0,3982 | 0,3980 | 0,3977 | 0,3973 |

# GLOSARIO

## AZAR

Casualidad. Caso fortuito. Fenómeno que no sigue una regla, un orden, una ley conocida. En Estadística se contrastan las probabilidades a favor de las hipótesis del investigador en cuanto al efecto de las variables independientes sobre las dependientes contra la probabilidad de que los resultados sean debidos al azar, a la pura casualidad.

## DISPERSIÓN

Característica de un grupo que nos informa del grado en que las puntuaciones de los integrantes de un grupo se sitúan de forma más o menos cercana la medida de posición de que se trate (por ejemplo, de la media aritmética). Un grupo en que todos su miembros obtienen una puntuación igual a la medida de posición tiene una dispersión de 0; sin embargo, no existe un valor fijo de dispersión máxima.

Las medidas de dispersión o variabilidad más importantes y utilizadas son la desviación típica o la varianza.

## CONTROL

El método científico pretende establecer relaciones causales entre las variables relacionadas en su hipótesis. Lograr una meta tan elevada como este exige del investigador el dominio de la situación, de forma que, teniendo bajo su dominio la variable independiente, controle el conjunto de circunstancias, hechos, personas... que, además de dicha variable, puedan influir en la dependiente.

Si no fuera así, quedaría la duda de si la relación encontrada se debe a la variable independiente, a alguna de esas otras variables –convertidas en extrañas, esto es, en hipótesis rivales a la suya- o la la interacción entre unas y otras.

## CORRELACIÓN

Entendemos por correlación la relación existente entre dos o más variables.

La correlación puede ser perfecta, positiva o negativa (valor de ± 1), nula (valor de 0), o imperfecta, que incluye toda la gama de valores que van de 0 a 1, tanto positivos como negativos. La correlación es positiva cuando los valores de las variables aumentan o disminuyen en la misma dirección, y negativa en caso contrario.

El índice de correlación –coeficiente de correlación- más conocido es el de Pearson, representado por $r_{xy}$ .

## DESVIACIÓN TÍPICA

Medida de dispersión o variabilidad. Estadísticamente es la raíz cuadrada de la media de la suma de las desviaciones individuales de un grupo de sujetos, elevadas al cuadrado, con respecto a la media aritmética de un grupo.

La varianza es un índice del grado en que las puntuaciones individuales se agrupan más o menos en torno a la media del grupo; si todas las puntuaciones individuales coincidieran con la media, la varianza sería 0; cuanto más se aparten de ella, mayor valor alcanzará la varianza.

Esta medida es muy importante en la Estadística inferencial ya que se utiliza en las pruebas de contraste de hipótesis.

## DIAGRAMA DE BARRAS

Representación gráfica especialmente adecuada a variables cualitativas; las barras, situadas unas a continuación de otras, tienen como base las diferentes categorías y como altura su frecuencia.

## ESCALAS DE MEDIDA

Al aplicar la regla de medida y la correspondiente unidad a un determinado objeto llegamos a un número. Pero los números resultantes no tienen todos las mismas propiedades ni, por tanto, se les pueden aplicar las mismas operaciones matemáticas.

Con los números más perfectos, propios de una escala de medida de cociente o de razón (edad, talla, peso) podemos utilizar todas las operaciones matemáticas. Con los de escalas de intervalo (temperatura), no, ya que no tienen un 0 absoluto. Hay números propios de una escala de medida ordinal, que admiten menos operaciones que los anteriores; ahora bien, dado que el orden tiene en alguna medida un carácter cuantitativo (por ejemplo: clase social) algunos autores clasifican, en ocasiones, a estas variables como cuasi-cuantitativas. Por último, hay números propios de escalas de medida nominal; aquí los números no indican cantidad sino diferencia: lo que es igual recibe el mismo número y lo que es diferente, un número distinto.

## ESTADÍSTICA

Ciencia que trata de analizar e interpretar los datos recogidos con algún propósito, como la investigación científica.

Algunos autores la definen afirmando que su objeto es el estudio de los fenómenos aleatorios; recuerde el lector que cuando hablamos de contrastar los efectos de diversas intervenciones lo que hacemos es asignar probabilidades a que tales efectos se deban al

puro azar (aleatoriedad) o a la intervención llevada a cabo por el investigador en condiciones de rigor o control de explicaciones alternativas.

Cuando trabajamos con los valores de las muestras la Estadística se denomina descriptiva; si de tales valores deseamos pasar a estimar los correspondientes a la población, la Estadística se conoce como inferencial; esta es más compleja pero es la que ofrece más utilidad u aplicaciones tanto al científico como al profesional.

La inferencia estadística pretende sacar conclusiones sobre gran número de datos a través de observaciones de parte de esos datos. Se trata de generalizar los datos de una muestra a la población de la que procede. Mediante la estadística inferencial se puede estimar parámetros y realizar contraste de hipótesis.

## ESTADÍSTICO

Valores obtenidos en una muestra. Los más conocidos son los agrupados bajo las medidas de posición o tendencia central (media, mediana*, moda*), las de dispersión* o variabilidad (desviación media, desviación típica, varianza) o los coeficientes de correlación. Suelen representarse con letras latinas ( x , s, r...).

A partir de ellos, por inferencia estadística, podemos estimar sus correspondientes parámetros con determinados niveles de probabilidad, asumiendo un riesgo de error tipo I prefijado por el investigador.

## ESTIMACIÓN DE PARÁMETROS

Se denomina así el procedimiento por el que se trata de estimar el valor de un estadístico, obtenido en una muestra, a toda la población de la que aquella forma parte.

Toda estimación asume un cierto margen de error, medido en términos de probabilidad; este error puede hacerse tan pequeño como desee el investigador, pero nunca podrá hablar en términos seguros, de certeza.

Al hablar en el texto del coeficiente de correlación nos hemos acercado al concepto y procedimiento de estimación de parámetros.

## EXPERIMENTO

Es la modalidad de investigación empírica* más exigente; como consecuencia, su aportación esencial es la posibilidad de establecer, con razonable seguridad, relaciones de causa a efecto entre una o varias variables independientes (v.i.) y otra denominada dependiente (v.d.).

Para poder lograrlo se deben cumplir determinadas exigencias: el investigador debe poder planificar la acción y provocar el fenómeno, ha de poder realizarlo en condiciones de control y debe contar con medidas de calidad, tan válidas, fiables y precisas como sea posible.

## HIPÓTESIS

Entendemos por hipótesis las conjeturas sobre la posible relación entre los elementos -variables- integrantes del problema. En los diseños experimentales se formulan hipótesis sobre la relación causal entre una o varias variables independientes (V.I.) y la variable dependiente (V.D.)

Una hipótesis se somete a prueba o se contrasta tratando de apreciar si las probabilidades a su

favor son sensiblemente superiores a una explicación por azar. Esta segunda hipótesis se

denomina nula y se representa por H 0 , frente a la del investigador (H 1 )

### HISTOGRAMA

Representación gráfica de las puntuaciones obtenidas por un conjunto de sujetos en una variable

cuantitativa. En el eje X se sitúan los límites de los intervalos; en el Y, la frecuencia del intervalo.

### INVESTIGACIÓN EMPÍRICA

Para Selltiz, “Investigar es buscar de nuevo, echar otra mirada más cuidadosa para averiguar más. Echamos otra mirada porque puede haber algo erróneo en lo que ya sabemos […]

La investigación científica ha de ser sistemática, organizada, disciplinada y rigurosa. Investigación empírica es aquella que acude a la experiencia, a los datos, para llegar a conclusiones en relación con las hipótesis de partida.

### MEDIA ARITMÉTICA

Medida de posición resultante de sumar todas las puntuaciones de un grupo y dividir el resultado por el número de integrantes del grupo, representado por N.

Su ventaja fundamental radica en que todas y cada una de las puntuaciones de la serie incluyen en su valor en forma proporcional al mismo. Es especialmente adecuada para niveles de medida de razón e intervalo.

### MEDIANA

Medida de posición resultante de ordenar las puntuaciones de mayor o menor, o viceversa, y encontrar la que ocupa el lugar central de la serie. Si la serie tiene un número par de casos, la mediana será la media de las dos centrales.

Su inconveniente fundamental es que en la mediana no influyen los valores de las puntuaciones sino solo el orden que ocupan. Dos series muy diferentes pueden tener la misma mediana.

## ESCALAS DE MEDIDA

Una medida, en sentido estricto, es el resultado de comparar una unidad con una cantidad. La cantidad “peso” la medimos comparándola con la unidad “Kilogramo” u otras mayores o menores. El resultado es el número.

La definición más amplia de “medida” se debe a Stevens: Medir es asignar numerales a los objetos o hechos de acuerdo con ciertas reglas. Un numeral puede ser un número o un símbolo, lo que permite admitir el nivel o escala de medida nominal.

En nuestros ámbitos, no siempre es tan fácil proceder a medir variables; la mayoría de las variables son construcciones o constructos elaborados por los científicos e investigadores, como en el caso de la inteligencia, el nivel de conocimientos, el autoconcepto, la tasa de inflación, el producto interior bruto o similares.

En tales casos, la medida consiste en la asignación de valores de acuerdo con ciertas reglas, como ocurre en una prueba objetiva, un cuestionario de actitudes hacia los inmigrantes, la tasa de mortalidad infantil, etc. Los números que resultan no tienen las mismas propiedades que en el caso del peso, de la talla o de la edad, números perfectos que permiten todo tipo de operaciones y que son propios de escalas de medida de razón o cociente.

Variables como la temperatura, perfectamente medibles, se diferencian de las anteriores en que el punto de partida –cero grados- no es fijo, además de poder presentar valores inferiores. Este tipo de variables forman parte de la escala de intervalos. Las que se limitan a indicar el orden en una serie (primero, segundo...) se ubican en las escalas ordinales; y en el caso de variables que no indican cantidad sino semejanza o diferencia (sexo, estado civil, clase social, grados universitarios...) la escala se conoce como nominal.

## MODA

También denominada Modo, es una medida de posición que coincide con el valor más repetido de la serie de valores.

Su inconveniente fundamental es que en aquellos valores menos repetidos que el de la Moda no cuentan para su obtención.

Resulta especialmente adecuada para el nivel de medida nominal.

## MODELO

Entendemos por “modelo” una representación simplificada de la realidad. Tal representación puede ser icónica, analógica, matemática.

Los modelos matemáticos tienen una gran utilidad en Estadística. En la medida en que unos datos empíricos sigan razonablemente un modelo, podemos aplicar las propiedades de este al tratamiento estadístico de aquellos.

En nuestro ámbito, modelo es, un tipo de distribución de datos teórico o ideal al que pueden tender distribuciones empíricas o reales de ciertas variables.

Por ejemplo: la variable motivación por los idiomas, una vez medida en un conjunto amplio de sujetos (muestra) puede acercarse o apartarse más o menos de un modelo ideal o teórico como es la denominada curva normal de probabilidades* o campana de Gauss.

Este modelo tiene unas propiedades; si nuestros datos medidos se acercan suficientemente al modelo, podemos aplicarles las propiedades del mismo, lo que nos permitirá analizar los datos y obtener conclusiones.

Para decidir si podemos considerar que unos datos empíricos se acercan suficientemente al modelo hasta hacerlos compatibles con él, disponemos de pruebas de bondad de ajuste, como es el caso de chi o ji cuadrado, cuyo símbolo es $\chi^2$ .

Este tipo de pruebas asignan una probabilidad a los datos empíricos sobre su acomodación o no al modelo, lo que permite al investigador aceptar o no la hipótesis de nulidad.

**MUESTRA. MUESTREO**

Entendemos por muestra un subconjunto de una población. La muestra debe ser representativa de la población, para lo que deberá contar con un tamaño suficiente y con una selección por procedimientos imparciales, como el muestreo aleatorio.

Muestreo es el procedimiento utilizado para seleccionar la muestra; el preferible es el denominado aleatorio simple.

**PARÁMETRO**

Entendemos por parámetro el valor de un determinado estadístico no en la muestra en que se obtiene sino en el total de la población. Si los estadísticos más comunes, como las medidas de posición y variabilidad (media: ; mediana: Md; moda*: Mo; desviación típica: s; varianza: $s^2$ ...) se suelen representar por letras latinas, los parámetros lo hacen por letras griegas ($\mu$ = media; $\sigma$ = desviación típica; $\sigma^2$ = varianza...).

**POBLACIÓN**

El término “población” se define como el conjunto de todos los casos o elementos que cumplen con las características que la definen: los varones, las mujeres, los estudiantes de Farmacia, los políticos, los abogados...

En ciencias sociales no suele estar muy claramente definida. El investigador desea generalizar los datos de la muestra a la población.

En los estudios empíricos no suele ser posible –ni, en la mayoría de los casos, aconsejable- estudiar todos los casos; se acude en su lugar a muestras, que deben ser representativas del conjunto total o población.

Por medio de la Estadística inferencial se pueden hacer estimaciones de los parámetros a partir de las muestras (por ejemplo: desde a μ)

**PROBABILIDAD**

Frente a los sucesos seguros se encuentran los probables. El tipo de seguros a las que es más adecuado aplicar la probabilidad es el de los fenómenos aleatorios.

Conociendo las diferentes manifestaciones de un fenómeno, como el número de caras de un dado o de los números de la lotería, podemos decidir la denominada probabilidad a priori, suponiendo, como debe ocurrir, que todas las caras del dado y todos los números tienen las mismas oportunidades. En el primer caso, la probabilidad de una cara cualquiera es de 1/6; en el segundo, suponiendo que tengamos 60.000 números, será de 1/60.000.

Para nosotros es importante conocer los modelos de probabilidad, como el de la curva normal.

Gracias a ella, a la regla matemática que la rige, podemos asignar probabilidades a los fenómenos que la siguen.

**SIGNIFICACIÓN ESTADÍSTICA**

Por lo general, todo investigador está interesado en saber si los valores obtenidos en una muestra, denominados estadísticos, representan a los de toda la población (parámetros).

A este procedimiento lo hemos denominado estimación de parámetros. Cuando el valor medido en una muestra representa al valor para toda la población afirmamos que ese estadístico es estadísticamente significativo. Si no fuera así, no podríamos considerar al citado estadístico como representante del parámetro: parámetro y estadístico serian valores de poblaciones diferentes.

Como hemos señalado, toda estimación asume un cierto margen de error, medido en términos de probabilidad; este error puede hacerse tan pequeño como desee el investigador, pero nunca podrá hablar en términos seguros, de certeza.

Algo similar podemos afirmar en los contrastes de hipótesis. Cuando un investigado plantea su hipótesis, por ejemplo: los resultados sobre el clima de aula –variable dependiente- serán mejores con un sistema A de disciplina que con otro B –variable independiente- (H 1 ) trata de mantener su hipótesis frente a una hipótesis alternativa –hipótesis nula o de nulidad, H 0.

Al final, después de aplica durante un tiempo los dos sistemas, llegará, por ejemplo, a dos medias aritméticas, y su problema será el de decidir si la diferencia entre ambas puede atribuirse a que el sistema A es mejor que el B o puede explicarse por casualidad, por azar (H0).

Si puede hacer lo primero, afirmará que las diferencias entre ambas medias aritméticas son reales, son estadísticamente significativas, y podrá mantener H 1 con una probabilidad a su favor

tan elevada como desee, pero nunca con certeza. En caso contrario, no podrá rechazar H 0 y tendrá que admitir que tales diferencias pueden ser explicadas por el azar.

## VALIDEZ

Utilizamos el término “validez” en dos contextos diferentes:

a) Como cualidad técnica de un instrumento de recogida de datos, indicando el grado en que tal instrumento mide lo que pretende y dice medir.

Como hemos indicado en el texto, dos manifestaciones de la validez, la concurrente y la predictiva, utilizan la correlación para poner de relieve la magnitud de la misma.

b) Como exigencia fundamental en los diseños de investigación experimental. La denominada validez interna, de darse, permite afirmar que los efectos medidos en la variable dependiente se deben a, y solo a, la variable independiente. Para ello el investigador debe controlar las variables extrañas. La validez externa se conoce como generalización, e informa del grado en que los resultados de la investigación pueden generalizarse.

## VARIABLES

Frente a una constante, la variable es aquella realidad que admite diversos valores, como la edad, la clase social, la inteligencia, el rendimiento académico o diferentes dimensiones o factores de la personalidad.

Cuando una variable solo admite valores enteros la denominamos discreta, tal como ocurre con el sexo, el estado civil, la clase social, o la carrera universitaria; las variables continuas pueden tener todo tipo de valores intermedios, como ocurre con la talla, el peso o la edad.

Las primeras pueden ser dicotómicas, si únicamente admiten dos valores o politómicas, en el caso contrario; en el primer caso se ha venido situando el sexo, mientras en el segundo podemos citar el estado civil.

Desde la perspectiva de la investigación las variables suelen clasificarse, en función del papel que desempeñan, en independientes, las manipuladas por el investigador, y dependientes, aquellas sobre las que se mide la influencia de las primeras; también podemos hablar de variables extrañas, esto es, variables que pueden convertirse en rivales de la hipótesis del investigador al influir sobre la dependiente junto a la independiente o en lugar de ella.

## VARIANZA

Medida de dispersión* o variabilidad. Estadísticamente es la media de la suma de las desviaciones individuales de un grupo de sujetos, elevadas al cuadrado, con respecto a la media aritmética de un grupo.

La varianza es un índice del grado en que las puntuaciones individuales se agrupan más o menos en torno a la media del grupo; si todas las puntuaciones individuales coincidieran con la media, la varianza sería 0; cuanto más se aparten de ella, mayor valor alcanzará la varianza.

Esta medida es muy importante en la Estadística inferencial ya que se utiliza en las pruebas de contraste de hipótesis; la más conocida e importante es la denominada F, o ANAVA (análisis de la varianza, aunque lo que se contrasta son medias aritméticas) que

atribuye a las diferencias entre medias una determinada probabilidad de que no sean explicables como consecuencia del azar. En muchos textos encontrará la expresión ANOVA (de analisys of variance) que se acomodan a ella.