Marco Polo Rivero R.
Abril 1976

4. ESTADISTICA APLICADA A LA HIDROLOGIA

de: Manual de Hidrologia
Tomo I
Rafael Heras

I-4. ESTADISTICA APLICADA A LA HIDROLOGIA

4.1. INTRODUCCION

4.1.1. Importancia de la estadística en la hidrología

Es práctica corriente utilizar la palabra estadística para designar los resultados de una serie de observaciones, presentados como conjuntos de números y ordenados generalmente en forma de cuadros, a los que pueden unirse otros números obtenidos mediante sencillas operaciones aritméticas, tales como frecuencias, valores medios, etc. Este aspecto de la descripción y resumen de masas de datos puede englobarse bajo la denominación de Estadística descriptiva, reservando la de Estadística matemática para la ciencia que, partiendo de la teoría matemática de probabilidades, tiene por objeto la obtención de conocimientos sobre conjuntos numéricos no conocidos, a partir de observaciones de conjuntos más reducidos.

Veamos cómo ambos conceptos encajan en el campo de la hidrología.

La hidrología de superficie constituye la parte más importante por corresponder a ella los ciclos más rápidos de circulación del agua.

Se engloban aquí gran parte de los ciclos subterráneos que tienen su origen o final en aguas superficiales y que corresponden a períodos de corta duración.

La hidrología es una ciencia que, más que muchas otras, está lejos de una madurez teórica que permita un apoyo decisivo en estudios de este tipo. Ello es debido a que. a pesar del gran número de estudios y teorías desarrollados, lo que pudiéramos llamar sus fuerzas motrices actuantes son de gran complejidad, tanto por su número como por las prácticamente ilimitadas maneras de actuar.

La mayoría de las causas que actuan en los ciclos hidrológicos superficiales son de carácter meteorológico y la propia meteorología se desarrolla fundamentalmente a través de la estadística, ya que es muy difícil llegar a un estudio matemático y preciso de los problemas físicos que condicionan los fenómenos hidrológicos.

Sin embargo, como los caudales de los ríos y sus cauces constituyen un complejo menos complicado y amplio que la atmósfera, es más fácil y viable estudiar estadísticamente los ríos a través de sus estaciones de aforos, al menos en los cursos principales y sus afluentes hasta un orden compatible con una red foronómica eficazmente simple. Por tanto, la meteorología y su estadística aplicada se utilizan para extrapolar hasta donde los aforos no pueden alcanzar, por tratarse de ríos pequeños donde no puede pretenderse que cada uno tenga sus instalaciones foronómicas propias, o para ampliar la extensión de las series, puesto que normalmente es más antigua la estadística meteorológica que la de aforos.

La recopilación de los datos hidrológicos y su ordenación estadística tienen como fin práctico su aplicación para dimensionar, con el mayor acierto posible, las obras que en el futuro han de utilizar los recursos hidráulicos (embalses, presas, captaciones, obras de conducción, centrales hidroeléctricas, etc.) y prever el régimen de explotación, de manera que se obtenga el mayor beneficio posible de las instalaciones construídas.

Se supone siempre que, en el futuro, el régimen hidrológico de un río tiene cierta relación con el pasado y se procura

obtener, del conocimiento de la estadística de caudales, referencias para prever, dentro de ciertos márgenes de seguridad, el régimen de caudales que pueda presentarse en el futuro.

Lo ideal sería que, de los datos estadísticos de caudales, pudieran deducirse leyes matemáticas que permitieran prever la sucesión futura de caudales en función de variables cronológicas o, al menos, de situaciones anteriores y algo lejanas. Un fenómeno que a los primeros observadores hubiera podido parecer complicado, como es la variación del nivel del mar en las costas por las mareas, ha podido ser estudiado en sus causas y, actualmente, se predicen con exactitud sus períodos y carreras.

No se ha conseguido aún encontrar las leyes que rigen los fenómenos meteorológicos que influyen fundamentalmente sobre el régimen hidrológico de los ríos, aunque se conozcan más o menos algunos de los factores que afectan a la hidrología, tales como el giro de la tierra, que influye sobre las direcciones de las corrientes de aire hacia los centros de mínimas presiones, el calentamiento de la tierra en sus diversas latitudes, la variación anual de temperaturas que influye, sobre la capacidad de retención de vapor de agua por el aire, sobre la fusión de nieves y hielos y sobre la evaporación.

Es conocida, también, la influencia de ciertos factores geográficos sobre las precipitaciones, tales como la distancia de los grandes mares, la existencia de cadenas montañosas y su configuración, etc. y es fácil estimar la influencia de los factores fisiográficos de una cuenca, sobre la proporción de volúmenes de agua precipitados, que se convierten en escorrentías y fluyen por los ríos, destacando entre las características de la cuenca que tienen mayor influencia la permeabilidad del terreno, pendiente de laderas y ríos, existencia de lagos, proporción de terrenos cultivados, naturaleza de los cultivos, etc.

La lluvia y otras precipitaciones se producen con carácter discontinuo, siendo muy difícil predecir con gran anticipación su cuantía, por lo que aparecen como fenómenos prácticamente aleatorios, aunque existan complicadas leyes que desconocemos y que pudieran explicar su irregularidad.

En resumen, el resultado que se desea obtener de los estudios estadísticos de meteorología es encontrar unas leyes que nos indiquen, con la mayor precisión posible, la influencia sobre los caudales de los factores meteorológicos y la dependencia de dichos caudales respecto a situaciones anteriores, pero, por no resultar exactas las leyes encontradas, se deben investigar, asímismo, las leyes de distribución aleatoria, que nos permitan tener idea de la dispersión de los posibles caudales respecto a estas leyes y la probabilidad con que se pueda esperar unas determinadas circunstancias hidrológicas.

4.1.2. Homogeneidad estadística

La naturaleza de la homogeneidad en el proceso hidrológico puede ser estudiada, estadísticamente, con respecto al tiempo y al espacio.

4.1.2.1. Homogeneidad del tiempo

Un proceso o series temporales pueden considerarse homogéneos en el tiempo si procesos idénticos, obtenidos en las series, resultan iguales. Así, la serie aleatoria pura y el proceso estacionario o series temporales son homogéneos en el tiempo. En hidrología, realmente, no hay registros homogéneos en el tiempo, pues existen varias clases de causas naturales o artificiales en los fenómenos hidrológicos que no hacen posible la homogeneidad.

Podemos establecer varios tipos de fenómenos que condicionan la homogeneidad del tiempo en las series hidrológicas. Su clasificación es la siguiente: tendencia, periodicidad y persistencia.

4.1.2.1.1. Tendencia

Es un crecimiento o decrecimiento sin dirección que varía respecto al valor medio de una variable hidrológica. Tal es el caso de la tendencia de las precipitaciones anuales en una estación.

Pueden utilizarse numerosas técnicas estadísticas para determinar la tendencia del fenómeno hidrológico. El método comunmente empleado es analizarla por el método de medias móviles.

En, prácticamente, todos los climas del mundo, la probabilidad de que se produzcan caudales más o menos altos no es uniforme a lo largo del año, sino que existen estaciones en que suelen ser más intensas las lluvias, o las temperaturas influyen manteniendo las precipitaciones en forma sólida durante el invierno y fundiéndolas en los meses más calurosos, o puede influir también la evaporación, en sentido opuesto, devolviendo a la atmósfera, durante los meses calurosos, parte del agua precipitada, influyendo también la transpiración de las plantas durante las épocas de actividad vegetativa.

En consecuencia, tendremos una evolución de los caudales medios, a lo largo del año, por lo que no sería admisible una ley de distribución de caudales sin tener en cuenta esta tendencia estacional.

4.1.2.1.2. Periodicidad

La estadística es una ciencia auxiliar que se utiliza, principalmente, para facilitar métodos para la recopilación, preparación, análisis e interpretación de datos numéricos, con objeto de conocer la estructura de los fenómenos de la masa de datos.

Los métodos estadísticos son, esencialmente, métodos para tratar datos obtenidos mediante operaciones reiterativas o susceptibles de repetición.

En el proceso de resolver un problema de la vida real por medio de la estadística pueden considerarse tres pasos: primero, elección de un modelo matemático; segundo, comprobación de la realidad del mismo; tercero, obtención de las conclusiones adecuadas de este modelo para resolver el problema propuesto.

El modelo matemático que un estadístico exige para una operación reiterativa es, generalmente, tal que le permita hacer predicciones sobre la frecuencia con que cabe esperar que -

se presenten ciertos resultados cuando la operación se repite un determinado número de veces.

Veamos esta periodicidad en los fenómenos hidrológicos. Ya en el ciclo hidrológico, propiamente dicho, aparece esta reiteratividad; fundamento del ciclo es que toda gota de agua, en cualquier momento que la consideremos, por ejemplo, en el momento en que es lluvia, recorre un circuito cerrado hasta volver a ser lluvia.

La gran variación interanual de las aportaciones de los ríos ha determinado que, desde hace mucho tiempo, se haya tratado de encontrar alguna periodicidad en la sucesión de aportaciones interanuales, y se ha tratado de correlacionar esta periodicidad con movimientos astronómicos o ciclos de actividad solar.

La indudable existencia de largas series de años más abundantes o más secos que la media, ha contribuído a alimentar las esperanzas de encontrar una cierta tendencia a la periodicidad en los regímenes meteorológicos e hidrológicos. Para estudiar la posible periodicidad de las lluvias y aportaciones anuales, se ha procurado examinar datos de gran duración, como los correspondientes a las crecidas del Nilo, las oscilaciones de nivel en algunos grandes lagos (especialmente el mar Caspio) y el espesor de las capas de crecimiento anual de los troncos de árboles. Se ha obtenido una periodicidad del orden de 35 años a partir de dichas observaciones, aunque también se han obtenido valores de 11 y 21 años en relación con las manchas solares.

Las observaciones de carácter local suelen, a veces, indicar períodos más cortos durante ciertas épocas recientes. Por ejemplo, sorprende que los años más secos durante los últimos 25 años, se hayan producido cada cuatro: 1945, 1949, 1953, 1957, 1961 (relativamente) y 1965, habiendo sido, en cambio, húmedos los años 1947, 1951, 1956 y 1960. Sin embargo, la periodicidad en largas series de datos anuales de aportaciones, no suele resistir un análisis estadístico, que no suele indicar dependencia muy estricta entre los caudales de cada año y los de n × p años antes, para cualquier valor que se dé a p como período de las oscilaciones.

Por consiguiente, es muy difícil detectar unas oscilaciones periódicas para las precipitaciones o aportaciones de la mayor parte de los ríos, lo que no quiere decir que la distribución interanual de aportaciones sea totalmente aleatoria, puesto que puede comprobarse la existencia de largos períodos de años en que los caudales medios son sensiblemente superiores o inferiores a los medios generales, siendo la dispersión de estas medias respecto a la media indefinida muy superior a la que correspondería calculando la dispersión de la media de dichas muestras.

Si se examinan largos períodos de tiempo se observa, sin embargo, que los caudales medios, aunque tengan persistencias más o menos marcadas, no presentan tendencias fuertes.

4.1.2.1.3. Persistencia

Debido a causas climatológicas y meteorológicas, se ha encontrado que los años lluviosos tienden a producirse en grupos y los años secos se producen también juntos. Esta tendencia de agruparse produce un efecto, como antecedente inmediato de las condiciones hidrológicas, que indica la persistencia en un fenómeno hidrológico.

Dado que este efecto juega una importante parte en el fenómeno hidrológico, él y, consecuentemente, la persistencia, están inversamente relacionados con el intervalo de tiempo entre las observaciones de tales fenómenos. Cuando el intervalo de tiempo es más corto, el efecto producido o la persistencia se vuelven más pronunciados. Como el efecto de la persistencia existe, el grado de aleatoriedad pura de los datos hidrológicos disminuye. La magnitud de la persistencia puede ser determinada por un análisis de correlación y depende del tipo de los datos hidrológicos.

Leopold describió la naturaleza de la persistencia, con referencia al análisis de la probabilidad, aplicada al problema del abastecimiento de agua, y Hurst, analizando los períodos más largos de datos de ríos en el mundo (1050 años de registros de niveles de agua del río Nilo en la estación pluviométrica de Ro da), llegó a la evidencia de que la tendencia en la ocurrencia de años lluviosos y años secos crece con la variabilidad de -- las medias de los períodos. En otras palabras, la variabilidad de grupos de datos pluviométricos en su orden natural de ocu-- rrencia es mayor que si se hubieran producido en una serie -- aleatoria.

Para ilustrar este punto, Leopold estudió la variabilidad de los valores medios de escorrentía fluvial para registros de varios períodos, según puede verse en la fig. I-4.1.

La curva punteada corresponde a datos agrupados tomados de largos registros de escorrentía fluvial de los Estados Unidos y Europa. Si la escorrentía fluvial anual se hubiese producido en una serie aleatoria, la variabilidad de las medias de los grupos disminuiría inversamente a la raíz cuadrada del número de años que comprende el grupo. Así, las medias de grupos de 100 años serían $1/\sqrt{100}$ ó -- 1/10, como variable de valores de 1 año. La curva definida representa esta secuencia de datos aleatorios. La diferencia entre la curva definida y la punteada representa el efecto de la persistencia (ver figura I-4.1.).

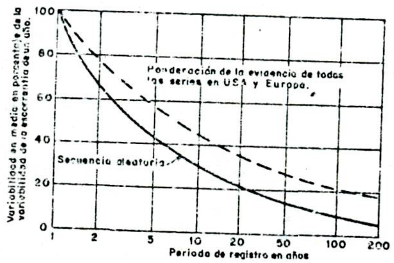


Fig I-4.1

La persistencia de los caudales en un río determinado es explicable, pues no puede haber una disminución brusca, ya que, aunque cesen las precipitaciones, el caudal del río seguirá alimentado por agua procedente de las capas freáticas y de los volúmenes de agua retenidos en los lagos y en los cauces de los ríos.

Es más difícil de explicar la persistencia de las precipitaciones, no sólo durante la misma estación del año, sino de un año al siguiente, aunque las estaciones lluviosas sean interrumpidas por otras secas. Se comprueba, en muchas, la existencia de dependencia masiva, tratando de obtener una correlación

entre las precipitaciones o aportaciones de cada año y el anterior, llegándose a índices de correlación significativos, que indican la dependencia.

4.1.2.2. <u>Homogeneidad del espacio</u>

La homogeneidad estadística meteorológica u homogeneidad estadística hidrológica en el espacio implica que, la ocurrencia de un fenómeno particular, meteorológico o hidrológico, producido dentro de todos los lugares llamados "áreas homogéneas" estadísticamente, está igualmente dentro de una diferencia estadísticamente tolerable. Debido a los cambios del medio geográfico, las áreas estadísticamente homogéneas son limitadas y pueden ser delineadas por análisis regional estadístico.

4.1.3. <u>Los datos hidrológicos considerados como variables aleatorias</u>

La aleatoriedad de un fenómeno está caracterizado por la propiedad de que su observación, bajo un conjunto de cir--- cunstancias análogas, no conduce siempre a los mismos resultados.

Las alturas de escala y los caudales que se observan en un tramo cualquiera de un curso de agua, las precipitaciones, etc., constituyen unos fenómenos en los cuales las variaciones cuantitativas dependen de factores comunes (fusión de la nieve, saturación del suelo, precipitación, condiciones del terreno, etc.). Pero nosotros no podemos expresar, por medio de una función, ni las relaciones recíprocas ni las relaciones con los fenómenos resultantes (altura de escala, precipitación). La relación entre tales fenómenos y los factores de los cuales de-- penden, no tienen un carácter determinado y únicamente se puede suponer. Se pueden designar tales fenómenos bajo el término de fenómenos aleatorios porque, cuando las condiciones adecuadas se reunen, su aparición no es cierta sino solamente previsible. Las condiciones para una crecida importante son, por --

ejemplo, una fusión brusca de la nieve en terreno empapado por precipitaciones importantes, pero puede ocurrir que todas estas condiciones, aun estando reunidas, no produzcan una crecida excesiva. La altura de escala de un día es más o menos independiente de la del día precedente, si bien puede esperarse que estos dos valores sean próximos, pero ni el sentido ni el valor del cambio dependen de la altura de agua, o al menos ellos dependen de manera preponderante de otros factores.

Los fenómenos absolutamente aleatorios, como los errores de medida, por ejemplo, pueden presentar unos valores cualquiera entre $+\infty$ y $-\infty$ pero con frecuencias diferentes; los valores límites son raros y las frecuencias aumentan a medida que se aproximan a la media aritmética de la serie de observaciones. Se ha verificado, teórica y prácticamente, el hecho de que cuanto más numerosas son las observaciones, la repartición de las frecuencias se aproxima más a la repartición llamada normal expresada por la fórmula

$$y = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{x^2}{2\sigma^2}}$$

En esta función x representa la dispersión, $(x = X - X_o)$, de un miembro cualquiera de la serie $X_1\ X_2\ \ldots\ldots\ X_n$ con relación a la media aritmética, $X_o = \frac{X_1 + X_2 + X_3 + \ldots + X_n}{n}$ para $n \to \infty$, e Y representa la frecuencia específica que corresponde a x. Se puede estudiar de la manera siguiente el sentido de la constante σ: por derivación, obtenemos la relación

$$\frac{dy}{dx} = \frac{-Xy}{\sigma^2}$$

de donde sacamos en conclusión, que el máximo de la frecuencia específica es para X = 0, y que el valor máximo es

$$Y_{max} = \frac{1}{\sigma\sqrt{2\pi}}$$

de donde

$$\sigma = \frac{1}{Y_{max}\sqrt{2\pi}}$$

Se ve claramente que Y_{max} es grande, es decir, la frecuencia de la media aritmética es alta y entonces la curva tendrá un máximo en el centro y ella estará próxima al eje de las x, lo que significa que los datos se agrupan alrededor de la media aritmética; por el contrario, si Y_{max} es pequeño, la curva de las frecuencias se aplasta, los datos no se agrupan alrededor de la media, y entonces son más dispersos. De donde podemos considerar que σ, que es inversamente proporcional a y, da el valor de la dispersión.

4.2. EL CONTRASTE DE DATOS

4.2.1. Planteamiento

4.2.1.1. Necesidad del contraste

Además de las imprecisiones relativas a la concepción - misma de los aparatos de medida, la explotación de las estaciones hidrológicas (pluviométricas, de aforos, etc.) da lugar a un cierto número de errores:

a) Errores de observación

- . lectura poco consciente o poco concienzuda
- . errores fortuitos de lectura
- . pérdidas en el aparato debidas al propio fenómeno (desbordamiento del pluviómetro, por ejemplo)
- . aparato ligeramente estropeado (no recoge el total del fenómeno)
- . errores debidos a la confluencia de varios fenómenos (en la lluvia, la evaporación)
- . mala colocación del aparato

b) errores de transcripción y de cálculo (los más frecuentes)

c) errores de copia

d) errores de impresión

e) errores de interpretación

Existe un pequeño porcentaje de datos falsos en las series cronológicas, que pueden ser debidos a las causas enunciadas anteriormente, y pueden clasificarse en dos tipos: accidentales y sistemáticos.

Es difícil descubrir los errores accidentales después de transcurrido algún tiempo, pues cualquier anomalía entre los datos correspondientes a diversas estaciones puede ser debida a algún error accidental, o bien a algún fenómeno meteorológico que afecte a la cuenca en que está situada la estación y no a las cuencas contiguas. Un error aislado, por otra parte, no puede afectar esencialmente a los estudios hidrológicos generales de un río por ejemplo, especialmente si su utilización está prevista regulando sus aportaciones mediante embalses de capacidad suficiente.

Mayor importancia tienen los errores sistemáticos pues, como consecuencia de ellos, los aforos (caudales, lluvias, temperatura, etc.) vienen incrementados o reducidos sistemáticamente, con lo que los resultados finales se desvían, pudiendo producirse grandes errores en los estudios de utilización y regulación que pudieran realizarse a partir de dichos datos.

Por todas las causas antedichas, es muy importante comparar los datos de las diversas estaciones con los registrados en otras de cuencas vecinas o con los de la misma cuenca, para poder detectar cualquier desviación sistemática.

4.2.1.2. Casos de aplicación más frecuente

4.2.1.2.1. Datos pluviométricos

Entre los errores sistemáticos se podrían considerar: los producidos por variaciones en las inmediaciones del pluviómetro (construcción de edificios, crecimiento de árboles, etc.) y algunos errores de anotación o suma.

El cambio de emplazamiento del pluviómetro puede considerarse como error sistemático, dado que normalmente a ambas estaciones se las designa por el mismo nombre, aunque en realidad es simplemente la desaparición de una estación y aparición simultánea de otra nueva. Si la situación de ambas es muy próxima, parece lógico conservar la misma denominación, siendo entonces preciso, para estudios cronológicos, homogeneizar la serie de observaciones, refiriendo los datos de una a los de la otra. Todo ello en el supuesto de que la instalación de ambas estacio-

nes se hubiera realizado correctamente, de acuerdo con las oportunas instrucciones.

En relación con la calidad de los datos podemos hacernos dos preguntas: ¿En qué condiciones y con qué garantía se puede afirmar que un dato es erróneo? y la segunda ¿Cuando es necesaria su corrección y en qué ocasiones basta con desecharlo?.

Para la primera pregunta parece oportuno decir, en relación con el porcentaje de datos erróneos, que éste es siempre muy pequeño.

La variabilidad de la lluvia con los efectos locales y más en especial la de tipo tormentoso, puede hacer que los datos destaquen claramente entre los de su zona, aún siendo verdaderos. Es más, esto puede ocurrir incluso de manera sistemática, siendo la estación verdaderamente representativa del punto en que está situada, pero no de su zona. Este sería el caso de una localidad en que la proximidad a una obra hidráulica la hiciera más expuesta a tormentas que otras de sus cercanías, pudiendo haberse instalado el pluviómetro, como verdaderamente representativo de la zona, con anterioridad a la construcción de la obra citada.

Al considerar intervalos de tiempo grandes, en general, se suavizan los efectos individuales de cada tormenta, haciéndose así más fácil el trazado de isoyetas. No obstante, puede ser de tanta importancia el efecto de estas tormentas, que puede reflejarse en el método de dobles acumulaciones, cambiando el gráfico de una recta a dos segmentos de recta paralelos, al efectuarse un salto en un cierto año en que la precipitación de una estación pueda ser debida, en su mayoría, a una o dos grandes tormentas.

Así pues, es aventurado considerar como error accidental un salto brusco en el gráfico de dobles acumulaciones, si no se consideran, en especial, la naturaleza de la precipitación en el período considerado y las características especiales de la zona en estudio.

En principio, cabría calificar de error el caso de que al confeccionar el gráfico de dobles acumulaciones, la estación en estudio experimentara un cambio brusco en sentido descendente. Pero, aún así, en la práctica es posible, ya que no suelen encontrarse varias estaciones base de comparación y al comparar con una sola es posible que esta última haya experimentado un gran incremento en la precipitación por efecto de una o dos grandes tormentas.

Así pues, no es fácil detectar rápidamente los datos erróneos y mucho menos los accidentales. Los sistemáticos se detectan mejor, debido a la repetición del error, sobre todo si se une a ello la experiencia en el trazado de isoyetas, en la recopilación y manipulación de los datos, el conocimiento del lugar o el sentido meteorológico.

Si se trata de dato mensual, antes de intentar corregirlo directamente, es preferible ver si se observa algún error mensual subsanable y en todo caso no utilizar un método de corrección de datos que no sea de aplicación al valor anual.

En relación con la segunda pregunta, podríamos decir lo siguiente:

Si se trata de una zona con adecuada densidad de estaciones, la supresión de una de ellas no significa error apreciable en el trazado de isoyetas en la zona. Si considerada la red pluviométrica adecuadamente distribuida, reducimos su densidad a la mitad, de modo que la red formada por las estaciones que quedan esté también bien distribuida, el incremento de error standard en el cálculo de la precipitación media de la zona es de un 5%.

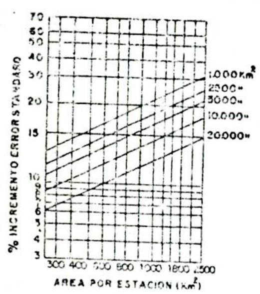


Fig I-4.2

En la figura I-4.2. puede verse ese efecto de la disminución de la densidad de la red en el error standard de la precipitación media calculada.

Si el dato erróneo corresponde a una zona con insuficiente densidad de la red pluviométrica, es mayor su necesidad, pero también es mayor la dificultad, tanto para poder afirmar que el dato es erróneo, como para proceder a su corrección. Esta, en caso necesario, deberá efectuarse prescindiendo del dato y utilizando uno de los sistemas de completado de datos.

4.2.1.2.2. Datos de aforos

Cuando se trata de aprovechar hidráulicamente algún río de escasa importancia o situado en regiones poco accesibles, es frecuente que se carezca casi totalmente de datos sobre las aportaciones de dicho río, no existiendo tiempo suficiente, antes de concluir el proyecto del aprovechamiento, para instalar una estación de aforos y obtener el registro de caudales durante un período de años suficientemente largo para conocer la evolución estacional de los caudales, las aportaciones medias anuales y la irregularidad de las aportaciones.

El que no se disponga de tiempo suficiente para reunir, antes de concluir el proyecto de un aprovechamiento hidráulico, todos los datos que se hubieran considerado convenientes, no debe excluir el comienzo de las observaciones hidrológicas y meteorológicas desde el momento en que se inicia el estudio, comenzando por estudiar la geomorfología de toda la cuenca para su comparación con otras cuencas análogas, debiendo realizar aforos aislados en diversas épocas del año e instalando, lo más rápidamente posible, estaciones pluviométricas y de aforos.

En otros ríos de mayor importancia, pueden existir datos de aforos en un cierto período de años, insuficientes para conocer las variaciones interanuales de aportaciones, bien por no haberse concedido importancia al posible aprovechamiento del río, o bien, por no haberse realizado correcciones frecuentes del tarado de la curva de caudales-alturas correspondiente a la estación de aforos, por lo que existen épocas en las que no se puede confiar en la exactitud de los datos.

Aún en las estaciones de aforos más cuidadas y más antiguas, los datos de aforos no se extienden a mucho más de 50 años, lo que es escaso para determinar la probabilidad de que se produzca un período extraordinariamente seco, con la recurrencia que requieren las garantías exigidas actualmente para la mayor parte de los aprovechamientos hidráulicos. La deficiencia de series de caudales suficientemente largas se hace especialmente sentir cuando se trata de dimensionar un aliviadero de presa, para el que se exige evacuar un caudal equivalente al máximo que se produciría con recurrencia de una vez cada quinientos años.

En estaciones de aforos antiguas, es frecuente la falta de datos de caudales, durante períodos más o menos largos, por las siguientes causas:

a) Por discontinuidad en las observaciones.

b) Por haberse averiado el limnígrafo o por haber sido llevada la escala por la corriente.

c) Por rebasar los niveles de agua la altura de escala o por no alcanzarla.

d) Por no existir, en ciertas épocas, curva de tarado de la sección de aforos.

e) Por lagunas de observación en ciertos períodos.

Actualmente, debido a la mecanización, es poco probable que se produzcan estas lagunas en los datos de aforos, por haberse perfeccionado la organización de observaciones, interpretación y publicación de datos hidrográficos. No es posible evitar que, en algún corto período, se produzca una interrupción en la toma de datos hidrográficos, pero es fácil reconstruir dichos datos, si no se deja transcurrir largo tiempo en el que se hayan borrado las señales y la memoria.

Vemos que, por unas u otras causas, es frecuente que los datos hidrográficos directos que se tienen para estudiar un río resulten insuficientes y deban ser completados con otros datos, menos precisos, estimados por procedimientos indirectos.

Cuando existan datos de alturas de escala, que no han sido transformados en caudales, por no existir entonces datos suficientes para representar la curva caudales-alturas o por no extenderse a los caudales de crecida por encima de una cierta altura, puede reconstruirse la relación de caudales, siempre que no se haya modificado el cauce en la sección de aforos por erosión o aterramientos. A estos fines se establecerá una curva con los caudales correspondientes a cada altura en los años próximos a aquéllos en que existen alturas de escala no transformadas en caudales. Con la curva correspondiente se calculan los caudales en los años en que faltan datos, comprobándose, por correlación con otras estaciones de aforos o por el método de las dobles acumulaciones, si los caudales calculados dan unas sumas anuales que no resulten absurdas. Cuando solo faltan los valores de caudales correspondientes a grandes alturas de escala, se puede prolongar la curva de caudales-alturas, bien sea como resultado de algunas observaciones posteriores, bien calculando los caudales en función de las alturas, por medio de alguna de las fórmulas aplicables a cauces abiertos y comprobando que, para los caudales aforados con los coeficientes de rugosidad y pendiente estimados, se alcanzan las alturas de escala observadas.

4.2.1.2.3. Caudales y precipitaciones

Hemos dicho que, generalmente, los datos sobre caudales de que se dispone para un estudio resultan insuficientes, por limitarse a un período de tiempo relativamente corto y, en algunos casos, porque son prácticamente inexistentes. Por otra parte, durante algunos períodos, pueden existir dudas sobre los datos de aforos y conviene contrastarlos con los estimados a partir de las precipitaciones en la cuenca.

En general, no hay demasiada abundancia de datos climatológicos, especialmente precipitaciones, tanto por lo que se refiere a la densidad de la red pluviométrica, como a la frecuencia de observaciones (faltan pluviógrafos) y a la extensión de las observaciones en el tiempo. Sin embargo, existen ciertas cuencas con suficientes pluviómetros instalados para poder estimar los volúmenes de agua precipitados sobre la cuenca, mientras no se dispone de datos de caudales. Por otra parte, suelen existir estaciones meteorológicas en las que las observaciones de precipitaciones se extienden a más de cien años, lo que permite prolongar la estadística de años secos y abundantes, aunque aplicando ciertas reservas a los datos, pues muchas de estas estaciones han cambiado de emplazamiento o bien han estado, en ciertas épocas, atendidas por personal inexperto.

Es preferible calcular las aportaciones de los ríos directamente, pero ante la escasez frecuente de datos, debe recurrirse a todos los procedimientos posibles para obtenerlos, por lo que es frecuente recurrir a la correlación con las precipitaciones para estimar las aportaciones.

Si se dispone de datos de precipitaciones y aportaciones anuales, durante suficientes años, en una o varias cuencas, cuyas condiciones fisiográficas y meteorológicas sean parecidas, puede intentarse establecer una ley de regresión entre precipitaciones y aportaciones.

a) Correlación lineal.

El caso más sencillo corresponde a la ley de regresión lineal:

$$\bar{A} = a\,\bar{P} - b$$

en regiones lluviosas, los coeficientes a y b no varían mucho de una cuenca a otra, siendo a ∿ 0,90 a 0,95 y b ∿ 400 ± 50 mm.

Para obtener, con cierta precisión, a y b, a partir de los datos disponibles, conviene calcular los valores medios de aportaciones y precipitaciones:

$$\bar{A} = \frac{\Sigma A}{n}$$

$$\bar{p} = \frac{\Sigma P}{n}$$

y los valores de varianzas y covarianzas:

$$\sigma_A^2 = \frac{\sum_1^n (A_i-\bar{A})^2}{n}$$

$$\sigma_p^2 = \frac{\sum_1^n (P_i-\bar{p})^2}{n}$$

$$\sigma_{Ap} = \frac{\sum_1^n (A_i-\bar{A})\,(P_i-\bar{p})}{n}$$

Si se desea obtener la mínima suma de desviaciones cuadráticas de las aportaciones, el valor del coeficiente a será:

$$a = \frac{\sigma_{Ap}}{\sigma_p^2}; \qquad b = a\,\bar{P}-\bar{A}$$

Si se desease calcular la recta de regresión de mínimos cuadrados de la distancia ortogonal, se emplearía el procedimiento indicado para la correlación entre caudales.

El coeficiente de correlación:

$$r = \frac{\sigma_{AP}}{\sqrt{\sigma_A^2 \times \sigma_p^2}}$$

da una idea de la bondad de la correlación y serviría para definir un término aleatorio que vendría a sumarse a los términos aP-b para indicar la probable desviación de las aportaciones, respecto a las calculadas por la regresión (fig. I-4.3.).

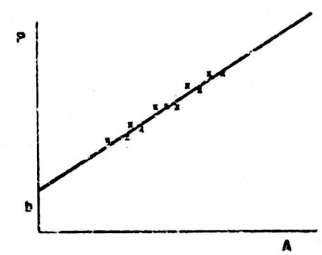


Fig.-I-4.3

La regresión lineal, en cuencas con precipitaciones abundantes, tiene la ventaja de ser válida, sin variar los coeficientes, para diversos años y para cuencas parciales de una misma cuenca general y, por ser lineal, es factible la superposición de las ecuaciones de varias cuencas para integrar una fórmula general.

Si la cuenca que se estudia tiene una gran reserva de agua en la capa freática al final de cada ciclo, las aportaciones de cada año pueden depender, no sólo de las de aquel año, sino también de las de años anteriores, en cuyo caso la ley de correlación tendría la forma:

$$A = a\ P_n + b\ P_{n-1} + \dots + m$$

pudiendo calcularse los coeficientes por los procedimientos explicados en el punto 4.2.4.

b) Correlaciones no lineales.

Si se aplican las leyes de correlación lineal a regiones secas, veríamos que en años extremadamente secos podrían obtenerse aportaciones negativas, lo que es un obstáculo para la aplicación de la regresión lineal en dichos casos.

En estas regiones secas puede ser más ventajoso, y muchos hidrólogos lo prefieren con carácter general, emplear fórmulas de correlación parabólicas de la forma:

$$A = \beta\ P^m$$

pudiendo calcularse el coeficiente β y el exponente m, de manera que sea mínima la suma de desviaciones cuadráticas o, gráficamente, representando los valores observados en papel logarítmico:

$$\log A = \log \beta + m \log P$$

Coutagne propone un valor para el exponente $m = 2$ y Becerril un valor $m = 3/2$, variando el coeficiente β según la precipitación y las características de la cuenca entre 0.005 y 0.030.

4.2.1.2.4. Otros casos

Cualquier fenómeno que experimente cambios, particularmente con el tiempo, y esté afectado de las características de periodicidad, tendencia, aleatoriedad, junto a la recurrencia y probabilidad, puede ser tratado como los fenómenos anteriores.

Así pues, como casi todos los fenómenos hidrológicos cambian con el tiempo, son aleatorios y, como la noción de probabilidad encaja bien, en su definición, podemos estimar que en sus datos iniciales hay una serie de errores que pueden detectarse por contraste entre los datos de base.

4.2.2. El método de las dobles acumulaciones

4.2.2.1. Bases

La teoría de la curva de dobles acumulaciones se basa en el hecho de que, representando en unos ejes coordenados las parejas de puntos definidas por las acumulaciones sucesivas de dos series de valores en el mismo período, la curva resultante es una línea recta, si los valores de las dos series son proporcionales; la pendiente de la recta representa la constante de proporcionalidad entre las dos series, o sea, que permanece constante sensiblemente la relación $\frac{\Sigma a_i}{\Sigma b_i} = K$, sin perjuicio de que la relación de dos valores puntuales homólogos no guarde, en general, la misma relación.

Si se produce un cambio en la pendiente de la curva de dobles acumulaciones, es que ha ocurrido una variación en la constante de proporcionalidad de las dos variables o que la proporcionalidad no es constante en el proceso acumulativo. Es posible que se desconozca la ley de variación de la relación entre las dos cantidades; sin embargo, el quiebro de la recta indica el momento en que ocurrió el cambio de la constante de proporcionalidad y la diferencia de pendientes en los dos tramos indica la diferencia de proporcionalidad. Estos cambios se pueden apreciar con mayor exactitud, si se escogen las escalas de abscisas y ordenadas de tal forma que la recta tenga una pendiente próxima a los 45°.

Para definir mejor la comparación, las acumulaciones de una de las variables se pueden comparar con las acumulaciones de un "grupo" de variables afines, para lo cual se calcula una variable tipo, media aritmética de las que forman el grupo, pues esta estación tipo está menos afectada por los errores que puedan existir en alguna de las variables del grupo. En general, y con objeto de poder utilizar las mismas comparaciones, se establecen éstas empezando por los datos más antiguos, puesto que de este modo se pueden seguir acumulando datos modernos en los futuros estudios. De todas formas, puede establecerse también - la comparación empezando por los últimos datos conocidos.

El número de estaciones que deben incluirse en la estación tipo, es función de la densidad de datos y, en teoría, se aconseja que no tenga menos de 10 estaciones. Si está formada - por menos de 10 estaciones, se debe comprobar la garantía de cada estación comparándola con la tipo, y eliminando sucesivamente aquellas estaciones con datos erróneos, aunque en la práctica la estación tipo puede tener hasta un mínimo de tres estaciones básicas.

Los cambios de pendiente formados por menos de 5 puntos no se consideran como representativos; en caso contrario, como norma general, indican errores de tipo sistemático.

4.2.2.2. Metodología operatoria

4.2.2.2.1. Aplicación a datos hidrológicos

Los métodos de aplicación de la técnica de la curva de dobles acumulaciones a los datos hidrológicos y la forma de -- utilizar los resultados, varían con el tipo de datos que son - analizados. En general, se utilizará este método para contraste de datos de precipitaciones y caudales y para estimar el valor absoluto del error producido por errores de tipo sistemático.

Vamos a analizar, sirviéndonos de guía las figuras --- I-4.4 y I-4.5, los casos más frecuentes y su interpretación, - suponiendo que la comparación la hacemos entre dos estaciones afines cuyos datos, en principio, suponemos tienen la misma garantía. Cuando una de ellas tenga mayor garantía o bien sea la media de una serie de estaciones, los errores que apreciemos - en la curva serán imputables a la otra estación.

A - En la serie de puntos encaja perfectamente una recta que, además, en este caso, pasa por el origen (en general - no tiene porqué pasar por el origen, pues la relación $\frac{ai}{bi}$ de -- los dos primeros valores de cada serie puede no coincidir con

$\frac{\Sigma ai}{\Sigma bi}$). Por tanto, en principio, los valores de las estaciones en todos y cada uno de los años son aceptables.

B - En la serie de puntos encaja perfectamente una recta que no pasa por el origen, lo que nos indica que existe proporcionalidad entre las dos estaciones en todos los años contrastados, excepto el primero. Si se tuviera que estimar la media de algún período en una estación, en función de la de la otra, se haría empleando la pendiente, pero los datos del primer año se deben respetar en ambas estaciones y contrastar posteriormente por comparación entre estaciones afines.

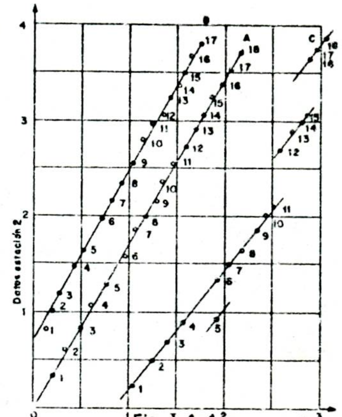


Fig I-4.4

C - Podemos encajar una serie de rectas paralelas a las que se ajustan todos los puntos; esto nos indica una proporcionalidad entre ambas estaciones, la correspondiente a la pendiente de la recta, aunque haya algunos años que en alguna de las estaciones hayan medido por defecto o por exceso o que estos años sean muy - irregulares en una de las estaciones. En la figura, estos años son 1, 5, 6, 12 y 16.

D - Cuando podemos encajar - dos rectas de diferente pendiente, es un caso típico de error sistemático. Para detectar cuál es la serie que tiene error, se establece - un proceso iterativo de comparación con otras series, y por la coincidencia o no coincidencia de cambios de pendiente se detecta la serie errónea. Más difícil es encontrar cuál es el período que registra bien, para lo cual se sigue también un proceso iterativo y se establecen comparaciones en valor medio suponiendo uno u otro período como bueno compa--

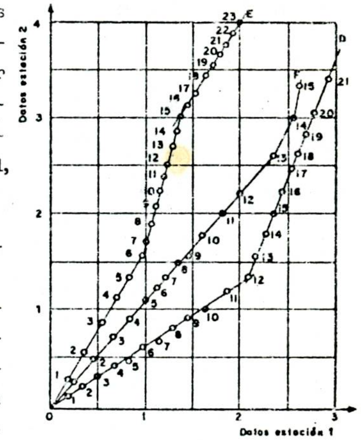


Fig I-4.5

rando la serie estudiada con otras series homólogas. Cuando el número de datos de ambos períodos es similar y no tenemos otra información, podemos aceptar como válido el período más moderno (esto no es científico, pero la experiencia demuestra que en el 80% de los casos el período más moderno es el más real, porque la estación está mejor vigilada o porque se le buscó un emplazamiento más adecuado, etc.).

En caso de error sistemático, es recomendable corregir la media del período estimado como erróneo e incluso, en caso necesario, estimar los valores anuales, aunque éstos no sean representativos y sea necesario contrastarlos posteriormente con las estaciones afines a la estudiada.

E - Sucede a veces que una estación (al igual que en el caso anterior, habría que comparar ambas con otras estaciones afines para saber cuál es la errónea) ha medido bien durante un período, ha tenido un error durante una serie de años y posteriormente la estación ha sido corregida. En el 95% de los casos, el período que midió el error es el central y hay que aplicar las correcciones explicadas en el caso anterior.

F - Cuando en una serie de años consecutivos los valores son superiores a la media del período considerado y creciente de año en año, la curva de dobles acumulaciones es una curva parabólica. En la figura, desde los años 12 a 15. Esta ley puede ser totalmente válida, puesto que la relación $\frac{\Sigma a_i}{\Sigma b_i}$ no se verifica para valores altos.

En todos los casos citados, excepto en los correspondientes a error sistemático, no es procedente corregir los datos que no encajen. Unicamente se hará por comparación directa con series de suficiente garantía; sin embargo, estos datos que no se ajustan en las dobles acumulaciones no debemos considerarlos en principio, pues para la utilización del método es lo mismo que el dato sea erróneo o irregular, ya que lo que se trata de obtener es la relación entre los valores medios acumulados, y los datos irregulares no son representativos. Como criterio general debe corregirse el menor número de datos posible y, sobre todo a escala puntual, la corrección por dobles acumulaciones no deberá hacerse, a no ser que se tenga suficiente información para saber con seguridad si él o los datos son erróneos.

4.2.2.2.2. Empleo del método en los estudios pluviómetricos

Constituye una técnica de análisis para ajustar las series de precipitación, con vistas a detectar y suprimir el efecto de factores no representativos, tales como cambios de emplazamiento del pluviómetro, cambio de tipo del mismo, del sistema de medida, etc. (todos ellos posibles de considerar como errores sistemáticos).

Esta teoría se basa en la comparación de la precipitación acumulada de una estación, con los valores acumulados del promedio de precipitación de varias estaciones básicas, y se efectúa representando ambas cantidades en unos ejes coordenados como abscisas y ordenadas.

La serie de puntos obtenida se ajustará tanto más a una recta, cuanto más constante sea la relación entre ambas cantidades que se acumulan. Si este gráfico dista mucho de parecerse a una línea recta, el procedimiento es inaplicable.

A la serie de puntos se le ajusta, de acuerdo con el gráfico de la figura 1-4.6, una o varias rectas. En el primer caso, la serie pluviométrica de la estación en estudio puede considerarse homogénea. En el segundo, puede transformarse en homogénea, refiriendo las observaciones de un tramo B a las del tramo A, mediante el producto de las precipitaciones correspondientes al primero por el cociente $\frac{P_a}{P_b}$ de las pendientes de los tramos A y B.

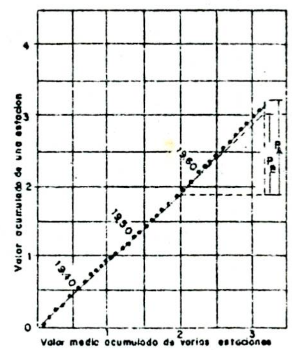


Fig-I-4.6

Primeramente, se deben situar sobre el mapa las estaciones pluviométricas, con una indicación del número de datos originales que poseen y una estimación de la media de estos datos. Para distribuirlas en grupos afines, se tienen en cuenta los siguientes criterios:

a) Los grupos deben tener de 3 a 10 estaciones

b) La media de las estaciones de cada grupo debe ser similar

c) Cada grupo debe tener, al menos, una estación con una serie de años suficientemente extensa (20 años como mínimo, aunque pueden utilizarse series de hasta 10 años)

d) La altitud de las estaciones del grupo debe ser similar

e) Las estaciones deben estar relativamente próximas

Los órdenes de magnitud aceptables para los puntos d) y e) dependen, fundamentalmente, de la densidad de la red, aunque en orden de magnitud, en general, la diferencia de altitud no debe ser mayor de 300 m. y la distancia no sobrepasar los 50 Km. Siempre que sea posible, el ideal es no sobrepasar el 50% de las cifras citadas.

Teóricamente, en cada grupo, y para aplicar correctamente el método de la curva de dobles acumulaciones, es conveniente comparar cada una de ellas con una estación "tipo" o "modelo", media aritmética de todas o parte de las estaciones de su grupo. De esta manera, los errores de las estaciones que constituyen el "modelo" se diluyen en éste (sobre todo los errores accidentales) y destacan más los de la estación que se compara.

En la práctica, principalmente por la no homogeneidad de las series, se suele considerar como estación tipo o modelo aquélla de serie más extensa y la más representativa a efectos, principalmente, de situación geográfica, siendo en este caso necesario contrastar la garantía de esta estación. Posteriormente, deben contrastarse los valores medios de todas las estaciones por correlación altitud-precipitación en zonas climatológicamente afines, que en general no coinciden con los núcleos considerados para establecer los grupos.

Una vez depuradas las estaciones, se establecería la comparación de la estación en estudio, con la media aritmética de las estaciones básicas con series homogéneas resultantes.

Los cambios de pendiente en la curva de dobles acumulaciones, pueden derivar de variaciones introducidas en el instrumento, en su sistema de medida o anotación de datos, o bien proceder de una causa física real. En los primeros casos, en general, el cambio de pendiente coincidirá con la introducción de las variaciones. Si no se observa esta coincidencia, se someterá la serie a una prueba estadística de valor significativo que, en caso de resultado afirmativo, conducirá a detectar el cambio introducido, o bien a la aceptación de una variación real en la pluviometría.

Como norma general, para poderse afirmar que existe un cambio de pendiente, cada uno de los tramos de recta debe afectar a cinco años como mínimo, para así poderse considerar fuera de la natural variabilidad de las lluvias.

Hay casos en que el cambio es debido a una variación real en la pluviometría de la estación en estudio. Un ejemplo de esto podría ser el incremento de precipitación de tipo tormentoso, en una localidad próxima al emplazamiento de una obra hidráulica de nueva construcción. Otro sería la variación temporal observada en una cierta estación, por efecto de una campaña de experimentación de lluvia artifical.

4.2.2.2.3. <u>Empleo del método en los estudios foronómicos</u>

Para contrastar los caudales registrados en las estaciones de aforos por el método de dobles acumulaciones, hay que realizar un estudio previo de la afinidad hidrológica de las cuencas receptoras de las estaciones; suponiendo que son análogas la geología, topografía, vegetación, etc., se establecen "a priori" las comparaciones por criterios de semejanza de superficie de cuenca y, posteriormente, partiendo de los datos físicos de cada cuenca (rectángulo equivalente, curva hipsométrica, índice de pendiente e índice de compacidad), se establecen las comparaciones definitivas en función, también, del número de datos de cada una de las series. En el ejemplo de la figura I-4.7. las comparaciones que se deben hacer son:

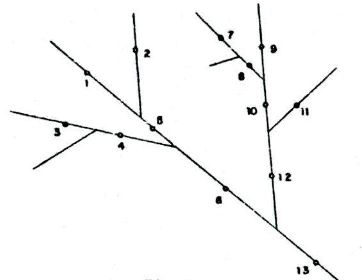


Fig-I-4.7

1 - 2	4 + 5 - 6	8 + 9 - 10
1 - 3	7 - 8	10 + 11 - 12
3 - 4	7 - 9	12 - 6
1 + 2 - 5	8 - 9	6 + 12 - 3
4 - 5	9 - 11	

El ajuste de dobles acumulaciones, en el caso de valores de caudales, dependerá fundamentalmente del grado de correlación de los mismos, siendo éste mucho mayor cuando se trata de estaciones en el mismo río, siendo importante también, para que el ajuste sea bueno, que la afinidad hidrológica entre las cuencas de las estaciones sea alta, y en función de ella se exigirá un ajuste más o menos afinado.

4.2.2.3. Alcance y limitaciones del método

Las limitaciones del método de dobles acumulaciones son las siguientes:

a) no es recomendable en zonas de gran altitud, aunque en la práctica se utiliza, contrastándole posteriormente con la correlación altitud-precipitación y anomalías

b) no se debe usar como base anual, en zonas donde el esquema de precipitación varía mucho de unas estaciones del año a otras

c) no es aconsejable para considerar la precipitación diaria o en períodos lluviosos individuales

Estas limitaciones, de tipo general, son importantes en nuestro país, dado que España es uno de los países más montañosos y que la latitud en que está situada la Península Ibérica, hace que la variación estacional de la precipitación sea verdaderamente considerable.

Aún aceptando que esto último no tiene, en general, más efecto que obligar a utilizar el método con base estacional y no anual, tampoco puede aconsejarse el procedimiento de estas condiciones o con base mensual, en caso en que la precipitación estacional o mensual tiene lugar en forma de un número pequeño de tormentas o períodos lluviosos individuales.

d) No da una indicación clara sobre la bondad de una estación como base de comparación con otra, estimándose esto de manera subjetiva, de acuerdo con la mayor o menor desviación de los puntos respecto a la recta que se interpola. Tampoco nos da una indicación en relación con el error del valor medio hallado.

4.2.3. Correlaciones gráficas

4.2.3.1. Introducción

La hidrología es principalmente una ciencia empírica, y una gran parte de los problemas con que se enfrenta el hidrólogo, presuponen un análisis de correlación o la aplicación de una relación derivada de dichos análisis.

Los métodos gráficos, ampliamente empleados hoy en hidrología, pueden establecer correlaciones entre caudales, aportaciones en diversos períodos de tiempo, caudales específicos, aportaciones específicas o alturas de escorrentía en toda la cuenca, así como entre precipitaciones, aguaceros, etc. Si llamamos (x_i y_i) los datos obtenidos en una información estadística, correspondiente al mismo tiempo, ($\bar{x}$, $\bar{y}$) las medias del conjunto de todos los valores, y representamos los puntos correspondientes a cada par de observaciones, veremos que, en general, existirá cierta tendencia a agruparse (figura I-4.8), pudiendo dibujarse a estima, una curva que representa la tendencia media de la nube de puntos (curva de regresión).

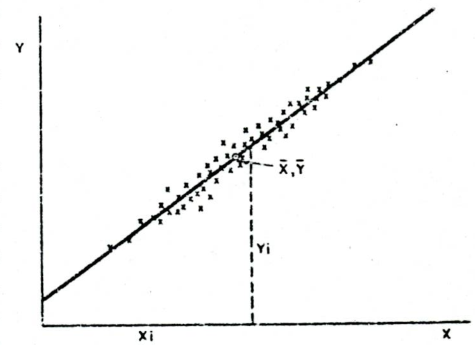


Fig.-I-4.8

Es por ello que representa no sólo un método estadístico, sino también un auxiliar inestimable de ayuda a las correlaciones analíticas que nos pueden indicar la bondad del método. Para su estudio hemos seguido el libro de Linsley, Kohler y Paulhus "Hidrología para ingenieros".

Se suele utilizar, para análisis y contraste de series, la comparación gráfica a escalas anual y mensual, que sirve fundamentalmente para detectar errores puntuales, estudiando gráficamente la representación cronológica de las series comparadas. También se suelen representar las curvas cronológicas de valores medios y las de valores máximos y mínimos para distintas probabilidades.

4.2.3.2. Correlación de dos variables

Si hay que emplear una relación lineal, la línea de regresión debe pasar por el punto definido por los valores medios observados ($\overline{X}$ e $\overline{Y}$) de las dos variables X e Y. Esto es cierto no sólo para la correlación gráfica, sino también para la línea de mínimos cuadrados. Fijando un punto ($\overline{X}$, $\overline{Y}$) sobre la recta, puede determinarse la pendiente, situando primeramente los datos (figura I-4.9) y determinando después las coordenadas medias de grupos de puntos. Los grupos seleccionados deben comprender todos los puntos que caen dentro de valores específicos de la variable independiente (X), prescindiendo del valor de Y, el factor a estimar. Si el total de puntos se divide en dos grupos, de aproximadamente el mismo número, la línea que une sus medias, pasará por la media de todos los puntos, siendo ésta la mejor línea que puede determinarse, con facilidad, gráficamente.

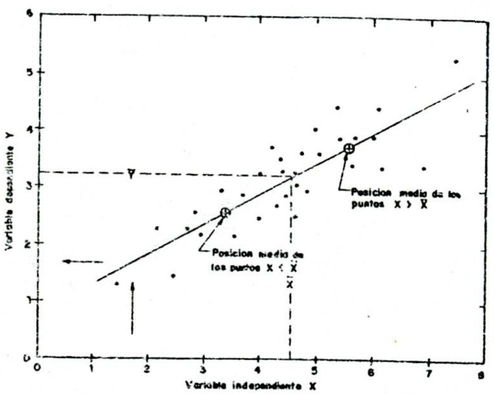


Fig-I-4.9

La relación determinada por medias de grupos tiene, generalmente, una pendiente ligeramente mayor (dy/dx) que la determinada por mínimos cuadrados. Al aumentar el grado de correlación, la diferencia entre las dos líneas disminuye, y en una correlación perfecta son coincidentes. La relación de valores medios tiende a minimizar la suma absoluta de las desviaciones, mientras que la relación de mínimos cuadrados, minimiza la suma de los cuadrados de las desviaciones.

Los valores medios de grupos pueden determinarse gráficamente, determinando primero las medias de dos puntos sucesivos (equidistantes de los puntos trazados). Los promedios de cuatro puntos son, pues, equidistantes de las medias de dos

puntos, etc. (figura I-4.10). Los puntos deberán agruparse con relación a valores de la variable independiente. De no ser perfecta la correlación resultará una línea diferente, si los puntos se agrupan de acuerdo con la variable dependiente, aumentando la diferencia al decrecer la correlación.

Si, al examinar los promedios de grupo (dos puntos, cuatro puntos, etc.) se decide que la relación es de forma curvilínea, podrá ajustarse una curva a las medias, con ayuda de una plantilla de curvas. La curva no pasa necesariamente por la media de todos los datos.

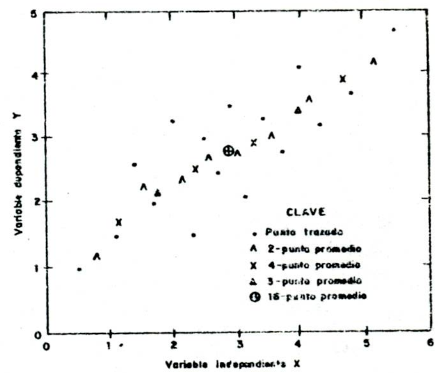


Fig.-I-4.10

4.2.3.3. Correlación de tres variables

Quizá el método más lógico de presentar una relación de tres variables sea por medio de un dibujo tridimensional (figu

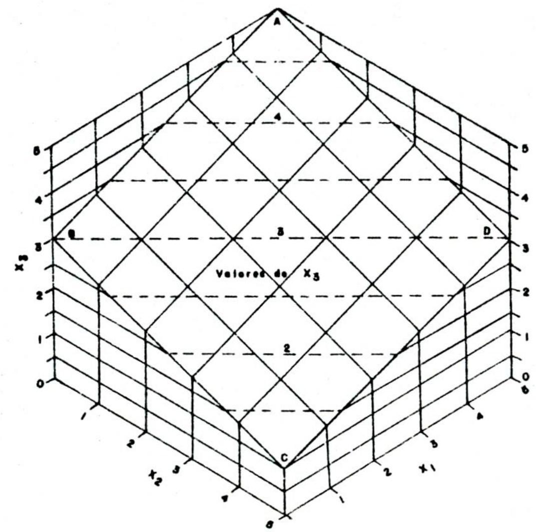


Fig. I-4·11

ra I-4.11). Se observará que las líneas que unen valores iguales de X_3 son rectas paralelas equidistantes, para incrementos iguales de X_3. Esto es cierto para cualquier superficie plana.

Dados los valores de dos cualesquiera de las variables, la tercera puede calcularse por medio del gráfico de la figura I-4.11. Un examen detallado de la figura demuestra que una vez construídas las tres familias de líneas sobre la superficie ABCD representada por la ecuación, las restantes partes del dibujo son ya superfluas. En la práctica, la superficie ABCD se proyecta sobre uno de los planos coordenados y se presenta como una familia de curvas sobre un papel cuadriculado.

Puede obtenerse un gráfico similar al ABCD de la figura I-4.11 por una serie de observaciones simultáneas de tres variables relacionadas. Representando los valores de Y en función de X_1 y acotando los puntos con los correspondientes valores de X_2, las líneas de igual valor de X_2 pueden ajustarse, en forma muy parecida a como se dibujan las curvas de nivel en la preparación de un mapa topográfico. En el análisis de correlación sin embargo, se pretende obtener una relación lógica entre las tres variables, por observaciones, sujetas a errores al azar. Más aún, la relación cierta envuelve frecuentemente valores adicionales. Por tales razones, no se intenta ajustar exactamente cada observación. Si el grado de correlación entre los datos trazados no es particularmente elevado, deberán determinarse los promedios de puntos agrupados con respecto a las variables independientes (X_1 y X_2), antes de intentar determinar las líneas de valor de X_2 constante (figura I-4.12). La situación de los puntos medios se determina gráficamente, rotulándose con el promedio de los valores X_2 del grupo.

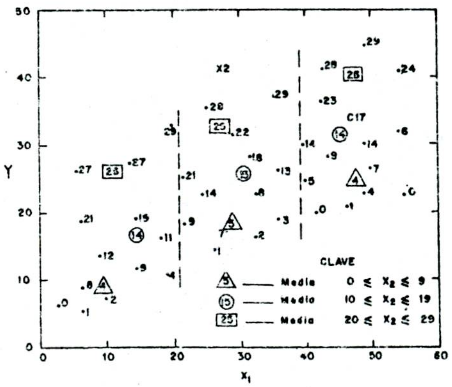


Fig I-4.12

Las discontinuidades bruscas en la separación o en la curvatura de las líneas no deben encuadrarse en una familia de curvas, basada en datos limitados, a no ser que puedan explicarse lógicamente. La complejidad de la familia de curvas obtenida debe estar de acuerdo, en un mínimo, con los datos disponibles y con el conocimiento previo de la relación. Una vez decidido el carácter general de la familia de curvas, el método de construcción (fig. I-4.13) es el siguiente:

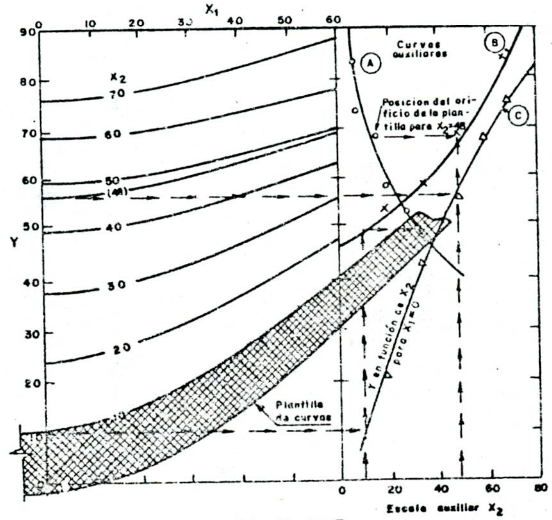


Fig. I-4·13

1. Determinar la posición óptima de la plantilla de curvas para diversas curvas X_2, anotando en cada caso la situación del orificio de la plantilla de curvas (u otro punto identificable cualquiera) y el valor de Y para algún valor fijo de X_1 (X_1 = 0).

2. Ajustar una curva (A) a los puntos definidos por las posiciones del orificio de la plantilla de curvas.

3. Volver a trazar los puntos empleados para definir la curva A, desviándolos horizontalmente hasta el valor apropiado de X_2 sobre la escala auxiliar (por ejemplo X_2 = 48 e y = 68) y ajustar la curva auxiliar B. Las curvas A y B fijan la posi-

ción del orificio de la plantilla de curvas para cualquier curva X_2 deseada.

4. Trazar los valores de Y (para $X_1 = 0$) obtenido en 1 - para los valores correspondientes de X_2 en la escala auxiliar y ajustar la curva auxiliar C. Esta curva fija la orientación de la plantilla de curvas para cualquier curva X_2 elegida.

5. Construir las curvas X_2 elegidas según la figura, para $X_2 = 10$. El orificio de la plantilla de curvas se coloca sobre la curva A en un valor de Y correspondiente a $X_2 = 10$ sobre la curva B. A continuación, se hace girar la plantilla para que el punto $X_1 = 0$ e Y, correspondientes a $X_2 = 10$, caiga sobre la curva C.

Si la familia de curvas ha de consistir en rectas paralelas equidistantes, la curva auxiliar C (también recta) determinaría la pendiente. Si hubiera que considerar separaciones variables y/o convergencia, en una familia de rectas, se requerirían dos curvas auxiliares análogas a la C (para los valores -- elegidos de X_1).

Todas las relaciones de tres variables pueden expresarse por un gráfico como el que acabamos de ver. En las funciones lineales, las líneas de valor constante están igualmente espaciadas y son rectas paralelas; en las funciones curvilíneas, pueden ser curvas espaciadas desigualmente, o ambas cosas a la vez. Líneas convergentes indican una función mixta, es decir, una -- función en la que el efecto de una variable independiente depende del valor de una segunda (por ejemplo $Y = X_1 X_2$).

4.2.3.4. Correlación coaxial

El método coaxial de correlación gráfica se basa en el - supuesto de que se omite cualquier factor importante de una relación; la dispersión de los puntos de un trazado de valores observados de la variable dependiente, en función de los valores calculados por la relación, tendrá su explicación, al menos parcial, en el factor omitido. En otras palabras, si los puntos de este trazado se acotan con los valores correspondientes al nuevo factor, podrá trazarse una familia de curvas para modificar

los valores calculados por la relación original. Se verá que - una relación coaxial es, en realidad, una serie de relaciones de tres variables dispuestas con ejes comunes para facilitar - el trazado y cálculo (figura I-4.14).

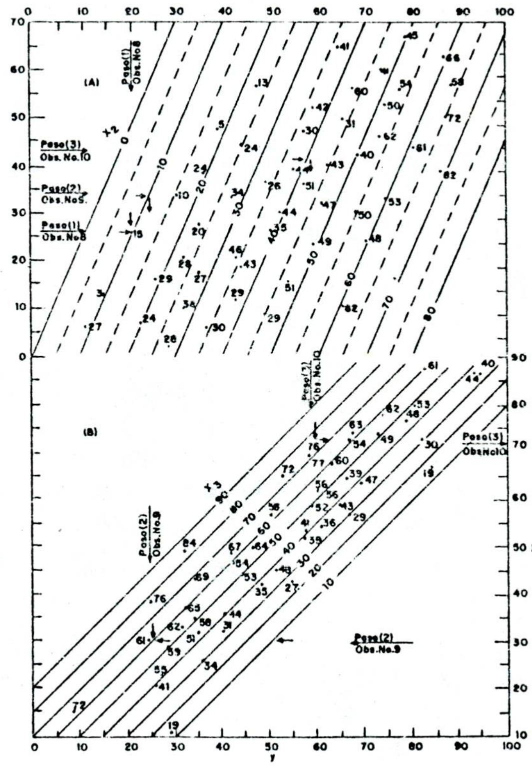


Fig I-4.14

A fines ilustrativos, supongamos que queremos desarrollar una relación para predecir Y partiendo de X_1, X_2 y X_3, y que los datos disponibles se han tabulado como indica la Tabla I. El análisis se realiza en la forma siguiente:

1. Trazar Y en función de X_1, para la serie completa de observaciones y marcar cada punto con el valor de X_2, como se indica en el gráfico A de la figura I-4.14. Ajustar una fami--

lia de curvas a los datos trazados (punto 4.2.3.3.). En este -- ejemplo resulta que las rectas paralelas se ajustan adecuadamente a los datos; sólo se emplearán funciones más complejas, cuando se note claramente o cuando puedan explicarse lógicamente.

2. A continuación se trazan los valores observados de Y en función de los calculados por el gráfico A (entrando con X_1 e interpolando dentro de las curvas X_2); se marca cada punto -- con el valor de X_3, como se indica en el gráfico B de la figura I-4.14 y se ajusta de nuevo una familia de curvas.

3. Entrando con X_1, X_2 y X_3 por este orden de sucesión - (figura I-4.14), calcular el valor de Y para cada observación - de la serie, tabulando el error correspondiente.

Si hubiera que introducir variables adicionales, se repetiría en esencia el punto 2 del proceso. En cada caso, calculada Y por la secuencia derivada del gráfico, se traza en función de los valores observados de Y, y se marcan los puntos con valores de la variable a introducir.

Como el gráfico A de la figura I-4.14 se dedujo sin te-- ner en cuenta los valores de X_3, la subsiguiente introducción - de este factor puede requerir una revisión del gráfico A, particularmente si hay correlación apreciable entre X_3 y X_1 o X_2. En otras palabras, la técnica coaxial requiere, generalmente, dos o más aproximaciones sucesivas para lograr los mejores resultados. Aunque se dispone de varias técnicas, sólo tratamos aquí - de un método para desarrollar la segunda y sucesivas aproxima-- ciones. Con este método el análisis continúa en la forma siguiente:

4. Reconstruir las curvas de la primera aproximación X_3, (figura I-4.14), sobre otra hoja de papel cuadriculado, como en el gráfico B de la figura I-4.15.

5. A continuación, en el gráfico A de la figura I-4.15, se trazan valores de X_1, en función de los valores de las coordenadas, obtenidos entrando en el gráfico B con los valores observados de Y (sobre el eje vertical) y X_3, marcando cada punto con X_2. Ajustar las curvas X_2 de segunda aproximación.

T A B L A I

Observación Nº	Y	X_1	X_2	X_3	Error absoluto: 1ª aproximación	Error absoluto: 2ª aproximación	Observación Nº	Y	X_1	X_2	X_3	Error absoluto: 1ª aproximación	Error absoluto: 2ª aproximación
1	88	58	58	61	1	0	27	44	45	24	53	2	1
2	32	21	26	51	3	1	28	78	68	45	62	2	2
3	74	34	53	63	0	0	29	50	27	35	64	4	2
4	73	47	62	30	1	1	30	36	39	24	44	2	0
5	86	40	82	22	1	0	31	54	16	51	41	0	0
6	29	2	28	59	4	0	32	64	66	41	47	4	4
7	35	18	27	58	4	1	33	61	33	47	56	3	3
8	21	26	15	41	1	3	34	62	41	43	56	1	1
9	30	34	10	61	0	2	35	65	11	62	39	4	1
10	68	43	40	77	4	4	36	68	31	50	60	0	0
11	57	48	30	58	2	2	37	55	40	44	36	1	1
12	67	57	60	19	2	2	38	35	28	20	62	3	1
13	87	64	66	44	3	2	39	59	53	42	43	2	3
14	49	8	29	84	0	3	40	73	61	41	54	4	3
15	65	51	31	72	1	1	41	32	10	36	31	1	3
16	70	25	48	76	2	1	42	39	48	5	76	1	0
17	15	13	3	72	5	2	43	74	54	50	49	1	1
18	52	31	44	39	0	1	44	45	19	43	48	5	4
19	80	45	61	53	2	2	45	43	21	46	27	0	0
20	26	16	29	34	2	1	46	49	37	26	67	2	0
21	37	6	30	65	3	1	47	87	52	72	40	2	2
22	11	6	27	19	3	1	48	42	34	34	35	1	2
23	23	7	24	55	7	2	49	47	57	18	54	2	1
24	59	24	49	52	1	1	50	43	12	29	69	1	2
25	57	37	51	29	1	0	Suma					101	71
26	77	57	54	48	0	0	Media					2.0	1.4

6. Trazar en el gráfico B (figura I-4.15.) los valores observados de Y en función de los calculados por el gráfico A, exactamente en igual forma que en el anterior número 2. Revisar las curvas X_3, como se indica, por los puntos trazados. En este ejemplo se observó que las curvas de primera aproximación, X_3, eran completamente satisfactorias.

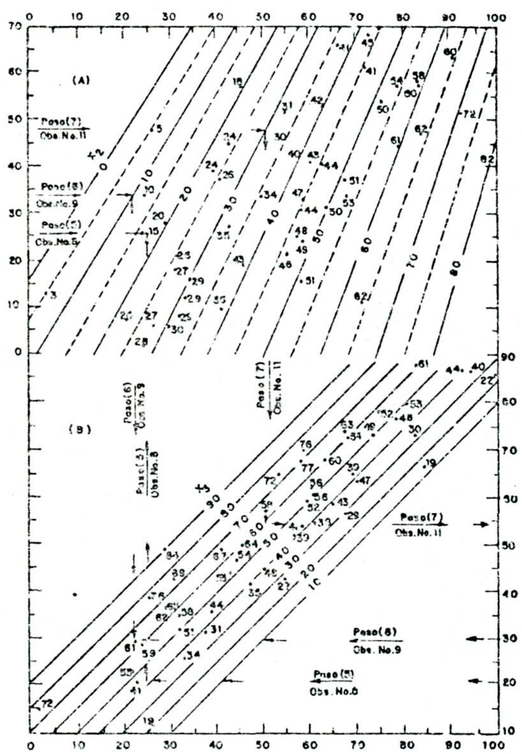


Fig.-I-4.15

7. Repetir el anterior número 3, empleando las curvas de segunda aproximación de la figura I-4.15., tabulando el error correspondiente (Tabla I). Si se realiza algún cambio en cualquiera de las familias de curvas, el error medio de la segunda aproximación suele ser apreciablemente inferior que el de la primera aproximación. Para evitar posibles perjuicios en la relación final, que podrían resultar de un ajuste inconsecuente de curvas, deberá calcularse la suma algebráica de los errores. Si dicha suma no es esencialmente cero, deberá corregirse una de las familias de curvas según se requiera.

Las aproximaciones tercera y subsiguientes, se obtienen por simple repetición de los números 4 al 7. Si el análisis implica cinco o más variables, habrá en la secuencia tres o más gráficos. En todo caso, el trazado de la segunda aproximación, para cualquier gráfico, se realiza entrando en la secuencia de gráficos, por ambos extremos, con los valores correspondientes de todos los factores (incluída la variable dependiente). El punto así determinado, se marca con el valor del factor para el que se desarrolla el gráfico, ajustándose una familia de curvas revisada.

Este método de correlaciones gráficas múltiples rinde, generalmente, buenos resultados en muchos problemas que implican funciones mixtas, tales como relaciones de lluvia-escorrentía y de niveles punta, donde existen tres o más factores independientes importantes. Si se conocen la forma aproximada y la separación de las familias de curvas, pueden dibujarse las curvas de primera aproximación para todas las variables "menos una" sin trazar los datos. Las curvas que representan la variable restante pueden desarrollarse a continuación trazándolas en la forma descrita. En general, se aconseja determinar las curvas del factor más importante por trazado. A continuación puede determinarse la posición definitiva de todas las familias de curvas por el proceso expuesto. Este método constituye la sustitución de un "efecto estimado" para cada variable con preferencia al supuesto efecto nulo de las variables introducidas subsiguientemente. Por lo tanto, queda reducido el número de aproximaciones requerido para lograr la solución final.

4.2.4. Correlaciones analíticas

4.2.4.1. Introducción

Existe cierta dependencia entre los caudales que circulan por diversos tramos de un mismo río, así como también entre los caudales que proceden de dos cuencas sometidas a análogo ré gimen meteorológico y cuyas cuencas tengan condiciones pareci-das de pendientes, red fluvial, permeabilidad, vegetación, etc. La existencia de la citada dependencia entre los caudales de dos cuencas, no quiere decir que exista una relación inequívoca y biunívoca entre caudales, ya que ciertos fenómenos meteorológicos o hidrológicos pueden afectar accidentalmente sólo a una de ellas, produciéndose cierta dispersión en los caudales de cada cuenca, correspondientes en el tiempo a los de la otra. Sin embargo, los métodos de la estadística permiten obtener una correlación entre los valores correspondientes a cada cuenca, consiguiéndose así una ampliación de los datos disponibles, con datos estimados que, si bien no serán exactos por no poderse tener en cuenta las causas de la dispersión, es probable que exista compensación de errores, con lo que los resultados del estudio con ellos efectuado no difieren grandemente con los que se hubieran obtenido en caso de disponer de datos directos.

Es preferible, para evitar errores, determinar, por procedimientos analíticos, las leyes de regresión, $y = f(x)$, que se acerquen más a los datos experimentales, para lo que se suele exigir que la suma de los cuadrados de las diferencias sea mínima.

Si consideramos las diferencias de ordenadas, elegiríamos una curva $y = f(x, a, b, c, ..)$ con diversos parámetros a, b, c, etc., que se determinarían por la condición de los mínimos cuadrados:

$$\frac{\partial \sum_1^n (y-y_i)^2}{\partial a} = 0; \quad \frac{\partial \sum_1^n (y-y_i)^2}{\partial b} = 0; \quad \frac{\partial \sum_1^n (y-y_i)^2}{\partial c} = 0$$

Los valores de la suma de cuadrados de los residuos

$$\chi^2 = \sum_1^n \frac{(y-y_i)^2}{y}$$

dará una idea de la precisión del ajuste.

La ley más sencilla de correlación es la regresión lineal:

$$y = a + b\,x$$

En algunos casos de correlaciones entre caudales de diversas cuencas o entre aportaciones y precipitaciones, no dá suficiente precisión la correlación lineal, siendo preciso recurrir a alguna ley de correlación con más parámetros, por ejemplo una correlación parabólica. Para determinar la curva de tarado (caudales-alturas) en una estación de aforos, no se podría emplear una recta.

En estos casos, la ley de regresión podría ponerse en la forma:

$$y = a + bx + cx^2 + \dots$$

determinando los coeficientes a, b, c, etc. de modo que $\Sigma(y_i - y)^2$ sea un mínimo respecto a ellos.

Es también posible, sin aumentar el número de parámetros, buscar leyes no lineales que se adapten mejor que las leyes de regresión lineal a la relación entre datos observados. Es posible transformar muchas de estas leyes en leyes lineales, mediante un cambio de variable y, después, operar gráfica o numéricamente, como si se tratase de investigar una recurrencia lineal.

Así:

$$y = a\,x^b$$

puede transformarse en:

$$\eta = \log y = \log a + b \log x = a_1 + b\,\xi$$

También:

$$y = a\,e^{bx}$$

$$\eta = \log y = \log a + bx \log e = a_1 + b_1\,\varepsilon$$

O bien:

$$y = \frac{a}{b + x}$$

en:

$$\frac{1}{y} = \eta = \frac{b}{a} + \frac{1}{a}x = a_1 + b_1\,\varepsilon$$

4.2.4.2. Correlaciones lineales

4.2.4.2.1. Planteamiento teórico

La ley más sencilla de correlación es la regresión lineal, $y = a\,x + b$.

Esta recta es tal que el momento de inercia de la masa de probabilidad del plano xy, con relación a esta recta, es mínimo. Ello nos está diciendo que el criterio establecido es que la suma de cuadrados de las distancias de los puntos de la nube a la curva sea mínima.

La condición precedente nos va a servir para determinar los parámetros a y b de la recta, sin más que obtener las derivadas parciales de la suma de cuadrados de las distancias de los puntos a la recta respecto a los dos parámetros e igualarlas a cero.

Ahora bien, podemos considerar tres casos, según que midamos las distancias de los puntos a la recta paralelamente al eje de las y, paralelamente al eje de las x o perpendicularmente a la recta. En los dos primeros casos aparecen dos rectas de regresión, según se tome como variable la x o la y, obteniendo las rectas de regresión lineal de y sobre x y de x sobre y. En el tercer caso obtenemos la recta de regresión ortogonal. Los momentos de inercia mínimos serán, respectivamente, I_y, I_x e I_o

Vamos a estudiar en este primer punto las dos primeras, dejando el punto siguiente para la tercera, de mayor importancia y aplicación práctica que las otras dos.

4.2.4.2.2. Parámetros utilizados

Valor medio

Sabemos que el valor medio o esperanza matemática de una variable aleatoria es el centro de gravedad de la masa de la distribución.

En la correlación lineal tendremos las medias (valor medio) de las variables x e y dadas por:

$$\bar{x} = \frac{1}{n}\sum_{1}^{n} x_i \qquad \bar{y} = \frac{1}{n}\sum_{1}^{n} y_i$$

Varianza y Covarianza

La varianza se define como el momento de inercia de la masa de la distribución, siendo la covarianza el momento central.

La varianza en x e y será:

$$\sigma_x^2 = \frac{\sum_{1}^{n}(x_i-\bar{x})^2}{n} \qquad \sigma_y^2 = \frac{\sum_{1}^{n}(y_i-\bar{y})^2}{n}$$

y la covarianza:

$$\sigma_{xy} = \frac{\sum_{1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n}$$

Desviación típica

La desviación típica o varianza de los residuos de los datos observados, respecto a la recta de regresión, será:

$$S_y = \sqrt{\frac{\sum_{1}^{n}(y_i-\bar{y})^2}{n}} = \sigma_y \sqrt{1-r^2}$$

$$S_x = \sigma_x \sqrt{1-r^2}$$

siendo r el coeficiente de regresión o correlación que, según veremos, es igual a:

$$\frac{\sigma_{xy}}{\sqrt{\sigma_x^2 \quad \sigma^?}}$$

4.2.4.2.3. Regresión lineal de y sobre x

Sea la ecuación de la recta buscada $y = a\,x + b$.

La distancia de un punto genérico, M (x_i, y_i), a la recta, -- $y = a x + b$, vale (figura I-4.16.):

$$\delta_i = ax_i + b - y_i$$

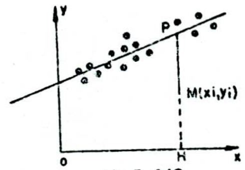


Fig. I-4.16

La varianza residual será:

$$S_y^2 = \frac{1}{N}\sum_1^n \delta_i^2 = \frac{1}{N}\sum_1^n (ax_i + b - y_i)^2$$

Para que sea mínima, debe cumplirse que:

$$\frac{\partial S_y^2}{\partial a} = \frac{2}{N}\sum_1^n (ax_i + b - y_i) \cdot x_i = 0$$

$$\frac{a}{N}\sum_1^n x_i^2 + \frac{b}{N}\sum_1^n x_i - \frac{1}{N}\sum_1^n x_i y_i = 0 \quad (1)$$

$$\frac{\partial S_y^2}{\partial b} = \frac{2}{N}\sum_1^n (ax_i + b - y_i) = 0$$

$$\frac{a}{N}\sum_1^n x_i + b - \frac{1}{N}\sum_1^n y_i = 0 \quad (2)$$

Sistema de ecuaciones, (1) y (2), de donde se obtienen - los valores de los parámetros a y b, y siendo:

$$\bar{x} = \frac{\sum_1^n x_i}{N}, \quad \bar{y} = \frac{\sum_1^n y_i}{N}, \text{ resulta:}$$

$$\left.\begin{array}{l} \frac{a}{N}\sum_1^n x_i^2 + b\bar{x} - \frac{1}{N}\sum_1^n x_i y_i = 0 \\ a\bar{x} + b - \bar{y} = 0 \end{array}\right\} \left.\begin{array}{l} a\sum_1^n x_i^2 + bN\bar{x} - \sum_1^n x_i y_i = 0 \\ a\bar{x} + b - \bar{y} = 0 \end{array}\right\} (3)$$

La recta de regresión lineal pasa por el centro de grave dad de los datos; si operamos en (3), nos queda:

$$b = \bar{y} - a\bar{x}$$

$$a\sum_1^n x_i^2 + \bar{x} N (\bar{y} - a\bar{x}) - \sum_1^n x_i y_i = 0 \Big\} \; a\left[\sum_1^n x_i^2 - N\bar{x}^2\right] = \sum_1^n x_i y_i - N\bar{x}\bar{y}$$

$$a = \frac{\sum_1^n \frac{x_i y_i}{N} - \bar{x}\bar{y}}{\sum_1^n \frac{x_i^2}{N} - \bar{x}^2} = \frac{\sum_1^n \frac{(x_i-\bar{x})(y_i-\bar{y})}{N}}{\sum_1^n \frac{(x_i-\bar{x})^2}{N}} = \frac{\sigma_{xy}}{\sigma_x^2};$$

en donde la varianza en x e y, valen:

$$\sigma_x^2 = \frac{\sum_1^n (x_i-\bar{x})^2}{n}; \quad \sigma_y^2 = \frac{\sum_1^n (y_i-\bar{y})^2}{n}; \quad \sigma_{xy} = \frac{\sum_1^n (x_i-\bar{x})(y_i-\bar{y})}{n}$$

Por lo tanto, la ecuación de la recta de regresión lineal de y sobre x queda de la forma:

$$y - \bar{y} = \frac{\sigma_{xy}}{\sigma_x^2}(x - \bar{x})$$

a su coeficiente angular se le llama coeficiente de regresión - de y sobre x y se representa por ρ_{yx}:

$$\rho_{yx} = \frac{\sigma_{xy}}{\sigma_x^2}$$

4.2.4.2.4. Regresión lineal de x sobre y

Si la recta de regresión buscada es $x = a y + b$, la distancia será $\delta_i = ay_i + b - x_i$, (figura I-4.17.) y la varianza - residual vale:

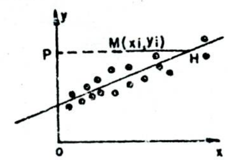


Fig. I-4.17

$$S_x^2 = \frac{1}{N}\sum_1^n \delta_i^2 = \frac{1}{N}\sum_1^n (ay_i + b - x_i)^2$$

Operando igual que anteriormente, obtenemos:

$$x - \bar{x} = \frac{\sigma_{xy}}{\sigma_y^2}(y - \bar{y})$$

y el valor del coeficiente de regresión será:

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_y^2}$$

4.2.4.3. Correlación ortogonal

4.2.4.3.1. Planteamiento teórico

La correlación ortogonal corresponde a uno de los ejes principales de inercia de la masa de la distribución. Por el teorema de Steiner sabemos que la recta tiene que pasar por el centro de gravedad de la masa. Tomaremos, pues, unos ejes que pasen por dicho centro de gravedad (figura I-4.18.).

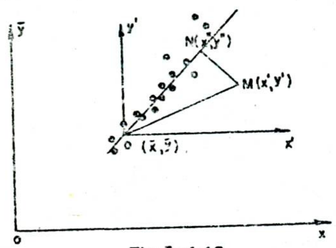


Fig I-4·18

Si α y β representan los cosenos directos de la recta, de ecuación:

$\frac{x}{\alpha} = \frac{y}{\beta}$, tendremos que $\alpha^2 + \beta^2 = 1$. (1)

En el triángulo rectángulo OMN, la distancia de un punto cualquiera, M, a la recta, vale:

$$\delta^2 = \overline{OM}^2 - \overline{ON}^2 = x'^2 + y'^2 - (\alpha x' + \beta y')^2$$

Si efectuamos la traslación de ejes, dividimos por n y lo extendemos a los distintos puntos, nos queda:

$$\Sigma \frac{\delta^2}{n} = \Sigma \frac{(x-\bar{x})^2}{n} + \Sigma \frac{(y-\bar{y})^2}{n} - \Sigma \frac{[\alpha(x-\bar{x}) + \beta(y-\bar{y})]^2}{n} \quad (2)$$

Vamos a obtener la recta tal que la distancia de cualquier punto a ella sea mínima, para lo cual aplicamos el método de los mínimos cuadrados. Por Lagrange, obtenemos el sistema:

$$\phi = \sigma_x^2 + \sigma_y^2 - \Sigma \frac{[\alpha(x-\bar{x}) + \beta(y-\bar{y})]^2}{n} + \lambda(\alpha^2 + \beta^2 - 1) \quad (3)$$

en donde sabemos que:

$$(4) \quad \sigma_x^2 = \Sigma \frac{(x-\bar{x})^2}{n}$$

$$(5) \quad \sigma_y^2 = \Sigma \frac{(y-\bar{y})^2}{n}$$

$$(6) \quad \sigma_{xy} = \Sigma \frac{(x-\bar{x})\,(y-\bar{y})}{n}$$

siendo $r = \frac{\sigma_{xy}}{\sqrt{\sigma_x^2 \cdot \sigma_y^2}}$, el coeficiente de correlación.

Si derivamos en (3) e igualamos a cero, obtenemos

$$\phi'_\alpha = 0 = -2\,\frac{\Sigma(x-\bar{x})[\alpha(x-\bar{x}) + \beta(y-\bar{y})]}{n} + 2\,\alpha\,\lambda$$

$$\phi'_\beta = 0 = -2\,\frac{\Sigma(y-\bar{y})[\alpha(x-\bar{x}) + \beta(y-\bar{y})]}{n} + 2\,\beta\,\lambda$$

Operando y teniendo en cuenta las expresiones (4), (5), (6), nos queda:

$$\left.\begin{array}{l} \alpha\,\sigma_x^2 + \beta\,\sigma_{xy} = \lambda\alpha \\ \alpha\,\sigma_{xy} + \beta\,\sigma_y^2 = \lambda\beta \end{array}\right\} \begin{array}{ll} \alpha(\lambda - \sigma_x^2) = \beta\sigma_{xy} & (7) \\ \alpha\,\sigma_{xy} = \beta(\lambda - \sigma_y^2) & (8) \end{array}$$

de donde:

$$\frac{\beta}{\alpha} = \frac{\lambda - \sigma_x^2}{\sigma_{xy}} = \frac{\sigma_{xy}}{\lambda - \sigma_y^2} \quad (9)$$

Operando en la expresión (3), y sustituyendo valores, nos queda:

$$\Sigma \frac{\delta^2}{n} = \sigma_x^2 + \sigma_y^2 - \alpha^2\sigma_x^2 + 2\,\alpha\beta\sigma_{xy} + \beta^2\sigma_y^2 = \sigma_x^2 + \sigma_y^2 -$$

$$- \alpha(\alpha\,\sigma_x^2 + \beta\sigma_{xy}) - \beta(\alpha\sigma_{xy} + \beta\sigma_y^2) = \sigma_x^2 + \sigma_y^2 -$$

$$- \lambda\,\alpha^2 - \lambda\,\beta^2 = \sigma_x^2 + \sigma_y^2 - \lambda,$$

en donde se han tenido en cuenta los resultados de las expresiones (7) y (8).

En definitiva, nos queda:

$$\Sigma \frac{\delta^2}{n} = \sigma_x^2 + \sigma_y^2 - \lambda \quad (10)$$

Si resolvemos el sistema formado por (7) y (8) nos queda:

$$(\lambda - \sigma_x^2)\,(\lambda - \sigma_y^2) - (\sigma_{xy})^2 = 0$$

$$(11) \quad \lambda^2 - (\sigma_x^2 + \sigma_y^2)\lambda + \sigma_x^2 \cdot \sigma_y^2 - (\sigma_{xy})^2 = 0,$$

ecuación de segundo grado en λ, que nos da dos raíces, λ_1 y λ_2, tal que $\lambda_2 > \lambda_1 > 0$.

De la expresión (11) sacamos:

$$\lambda_1 + \lambda_2 = \sigma_x^2 + \sigma_y^2$$

$$\lambda_1 \cdot \lambda_2 = \sigma_x^2 \cdot \sigma_y^2 - (\sigma_{xy})^2$$

El coeficiente angular de la recta de regresión ortogonal vale:

$$m = \frac{\beta}{\alpha} = \frac{\lambda_2 - \sigma_x^2}{\sigma_{xy}} = \frac{\sigma_{xy}}{\lambda_2 - \sigma_y^2}$$

La desviación típica vale

$$\sqrt{\Sigma \frac{\delta^2}{n}} = \sqrt{\lambda_1}$$

Obteniendo como ecuación de la recta de regresión ortogonal:

$$y - \overline{y} = m\,(x - \overline{x})$$

Suponiendo que la distribución de frecuencias de las desviaciones del conjunto de los puntos respecto a la recta de regresión ortogonal fuese normal, en el interior de la banda obtenida, trazando paralelas a la recta de regresión a una distancia $2\sqrt{\lambda_1}$, estaría comprendido el 95% de los puntos de la nube (figura I-4.19.). Así pues, aunque la hipótesis de distribución normal no es muy rigurosa, el dibujo de la banda en el gráfico de la correlación da una idea de la calidad de ésta y permite estimar, en una primera aproximación, los valores no ajustados.

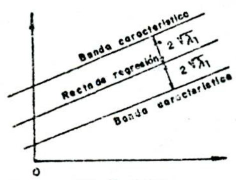


Fig.I-4.19

Puede suceder que $\lambda_1 = \lambda_2$. En este caso sabemos que cualquier recta que pase por el centro de gravedad de la masa es eje principal de inercia, habiéndose transformado la elipse de inercia en una circunferencia. El problema tiene entonces infinitas soluciones.

4.2.4.3.2. Coeficiente de correlación

A la media geométrica de los coeficientes de regresión, ρ_y y ρ_x, antes hallados, se le llama coeficiente de correlación de las dos variables y se representa por ρ ó r. Su expresión será

$$r = \rho = \sqrt{\rho_x\,\rho_y} = \frac{\sigma_{xy}}{\sqrt{\sigma_x^2\ \sigma_y^2}}$$

Los valores de ρ tienen que estar comprendidos en el intervalo (-1, +1). Además, cuanto más se aproxima $|\rho|$ a la unidad, menor será el valor de los momentos de inercia, I_y e I_x. En particular, para $|\rho| = 1$ resulta

$$I_x = I_y = 0$$

lo que quiere decir que toda la masa se encuentra sobre una recta, en la que se han confundido también las dos rectas de regresión.

Por ello, ρ mide, en cierto modo, la bondad del ajuste de los puntos a una recta.

En el caso de la regresión ortogonal, cuando $|\rho| = 1$, será

$$\sigma_{xy} = \sqrt{\sigma_x^2\,\sigma_y^2}$$

y la ecuación en λ se convierte en

$$\lambda^2 - \lambda\,(\sigma_x^2 + \sigma_y^2) = 0$$

cuya mayor raíz es

$$\lambda_1 = \sigma_x^2 + \sigma_y^2$$

con lo que resulta también

$$I_o = 0$$

y todos los puntos están sobre la misma recta.

Vemos que las dos rectas de regresión no coinciden en sus coeficientes angulares (fig. I-4.20), a no ser que sea:

$$\frac{(\sigma_{xy})^2}{\sigma_x^2 \cdot \sigma_y^2} = 1$$

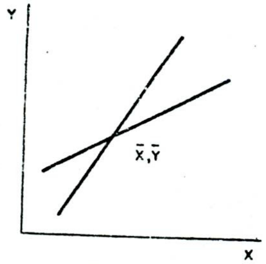


Fig.-J- 4.20

en cuyo caso todos los valores de los puntos (x_i, y_i) quedarían situados sobre una recta.

Si $r = 1$, no existirá ninguna dispersión en los residuos y, por consiguiente, todos los puntos quedarán situados en la recta de regresión, que será única para x e y.

Para valores pequeños de r la correlación lineal no tiene sentido, lo que no quiere decir que x e y sean independientes (correlación circular). Si $r = 0$, la dependencia entre las dos variables no existe.

En caso de ser $r < 0$, indica la existencia de anticorrelación.

El coeficiente de correlación se utiliza para determinar el grado de dependencia lineal que existe entre las dos variables.

4.2.4.4. Correlaciones no lineales

4.2.4.4.1. Correlación parabólica

Vamos a calcular la regresión parabólica de y sobre x.

Si la ecuación de la parábola es $y = a_o + a_1 x + a_2 x^2$, la distancia de un punto genérico, $M (x_i y_i)$, a la parábola, medida en dirección paralela al eje de las y, vale:

$$\delta_i = a_o + a_1 x_i + a_2 x_i^2 - y_i .$$

La varianza residual vale:

$$S_y^2 = \frac{1}{N} \Sigma \delta_i^2 = \frac{1}{N} \Sigma (a_o + a_1 x_i + a_2 x_i^2 - y_i)^2 .$$

Haciendo mínima la expresión de $\Sigma \delta_i^2$, tenemos:

$$\frac{\partial S^2 y}{\partial a_0} = 0 = 2 \cdot \frac{1}{N} \Sigma (a_0 + a_1 x_i + a_2 x_i^2 - y_i)$$

$$\frac{\partial S^2 y}{\partial a_1} = 0 = 2 \cdot \frac{1}{N} \Sigma (a_0 + a_1 x_i + a_2 x_i^2 - y_i) \cdot x_i$$

$$\frac{\partial S^2 y}{\partial a_2} = 0 = 2 \cdot \frac{1}{N} \Sigma (a_0 + a_1 x_i + a_2 x_i^2 - y_i) \cdot x_i^2$$

de donde operando:

$$\left.\begin{array}{l} a_0 N + a_1 \Sigma x_i + a_2 \Sigma x_i^2 = \Sigma y_i \\ a_0 \Sigma x_i + a_1 \Sigma x_i^2 + a_2 \Sigma x_i^3 = \Sigma x_i y_i \\ a_0 \Sigma x_i^2 + a_1 \Sigma x_i^3 + a_2 \Sigma x_i^4 = \Sigma x_i^2 y_i \end{array}\right\}$$

Sistema de tres ecuaciones con tres incógnitas a_0, a_1, a_2, que resuelto nos dan como valores:

$$a_o = \frac{\begin{vmatrix} \Sigma y_i & \Sigma x_i & \Sigma x_i^2 \\ \Sigma x_i y_i & \Sigma x_i^2 & \Sigma x_i^3 \\ \Sigma x_i^2 y_i & \Sigma x_i^3 & \Sigma x_i^4 \end{vmatrix}}{\begin{vmatrix} N & \Sigma x_i & \Sigma x_i^2 \\ \Sigma x_i & \Sigma x_i^2 & \Sigma x_i^3 \\ \Sigma x_i^2 & \Sigma x_i^3 & \Sigma x_i^4 \end{vmatrix}} = \frac{\Sigma y_i \left[\Sigma x_i^2 \Sigma x_i^4 - (\Sigma x_i^3)^2\right] + \Sigma x_i y_i \left[\Sigma x_i^3 \Sigma x_i^2 - \Sigma x_i \Sigma x_i^4\right] + \Sigma x_i^2 y_i \left[\Sigma x_i \Sigma x_i^3 - (\Sigma x_i^2)^2\right]}{N \Sigma x_i^2 \Sigma x_i^4 + 2 \Sigma x_i \Sigma x_i^2 \Sigma x_i^3 - (\Sigma x_i)^2 \Sigma x_i^4 - N (\Sigma x_i^3)^2 - (\Sigma x_i^2)^3}$$

$$\frac{\begin{vmatrix} N & \Sigma y_i & \Sigma x_i^2 \\ \Sigma x_i & \Sigma x_i y_i & \Sigma x_i^3 \\ \Sigma x_i^2 & \Sigma x_i^2 y_i & \Sigma x_i^4 \end{vmatrix}}{\begin{vmatrix} N & \Sigma x_i & \Sigma x_i^2 \\ \Sigma x_i & \Sigma x_i^2 & \Sigma x_i^3 \\ \Sigma x_i^2 & \Sigma x_i^3 & \Sigma x_i^4 \end{vmatrix}} = \frac{\Sigma x_i y_i \left[N \Sigma x_i^4 - (\Sigma x_i^2)^2\right] + \Sigma x_i^2 y_i \left[\Sigma x_i \Sigma x_i^2 - N \Sigma x_i^3\right] + \Sigma y_i \left[\Sigma x_i^2 \Sigma x_i^3 - \Sigma x_i \Sigma x_i^4\right]}{N \Sigma x_i^2 \Sigma x_i^4 + 2 \Sigma x_i \Sigma x_i^2 \Sigma x_i^3 - (\Sigma x_i)^2 \Sigma x_i^4 - N (\Sigma x_i^3)^2 - (\Sigma x_i^2)^3}$$

$$\frac{\begin{vmatrix} N & \Sigma x_i & \Sigma y_i \\ \Sigma x_i & \Sigma x_i^2 & \Sigma x_i y_i \\ \Sigma x_i^2 & \Sigma x_i^3 & \Sigma x_i^2 y_i \end{vmatrix}}{\begin{vmatrix} N & \Sigma x_i & \Sigma x_i^2 \\ \Sigma x_i & \Sigma x_i^2 & \Sigma x_i^3 \\ \Sigma x_i^2 & \Sigma x_i^3 & \Sigma x_i^4 \end{vmatrix}} = \frac{\Sigma x_i y_i \left[\Sigma x_i \Sigma x_i^2 - N \Sigma x_i^3\right] + \Sigma x_i^2 y_i \left[N \Sigma x_i^2 - (\Sigma x_i)^2\right] + \Sigma y_j \left[\Sigma x_i \Sigma x_i^3 - (\Sigma x_i^2)^2\right]}{N \Sigma x_i^2 \Sigma x_i^4 + 2 \Sigma x_i \Sigma x_i^2 \Sigma x_i^3 - (\Sigma x_i)^2 \Sigma x_i^4 - N (\Sigma x_i^3)^2 - (\Sigma x_i^2)^3}$$

4.2.4.4.2. Regresión parabólica por polinomios ortogonales

Sean dos variables, xy, que toman una serie de valores - x_1y_1, $x_2y_2 \ldots x_ny_n$. Sea n_i el número de veces que aparece la pareja x_iy_i. Supondremos que todas las n_i son la unidad pues si no lo fuera pondríamos varias veces x_jy_j según el valor de n_j. Si se representan los puntos en el plano xy se obtiene una nube de puntos. Vamos a tratar de adaptar a la nube una parábola del tipo

$$y = a_0 + a_1x + \ldots + a_hx^h$$

sea

$$P_h(x) = C_{h0} + C_{h1}\,x + C_{h2}\,x^2 + \ldots C_{hh-1}\,x^{h-1} + c_{hh}\,x^h$$

un polinomio cualquiera de orden h. Los polinomios son ortogonales (asociados a los valores $x_1\ x_2 \ldots x_n$) cuando

$$\sum_{i=1}^{n} P_h(x_i)\,P_j(x_i) = 0 \quad (h \neq j)$$

siendo las x_i diversos valores particulares de x.

Vamos a considerar las parábolas

$$y = b_0\,P_0(x) + b_1P_1(x) + \ldots + b_hP_h(x) + \ldots + b_kP_k(x)$$

siendo los polinomios ortogonales dos a dos.

El criterio que vamos a seguir es que la suma de cuadrados de distancias de los puntos de la nube a la curva, sea mínima. Tendremos:

$$E = \sum_{i=1}^{n} \{y_i - b_0P_0(x_i) - b_1P_1(x_i) \ldots -b_hP_h(x_i) \ldots -b_kP_k(x_i)\}^2$$

Derivando respecto a b_h

$$0 = \frac{\partial E}{\partial b_h} = \sum_{i=1}^{n} \{y_i - b_0P_0(x_i) - b_1P_1(x_i) - \ldots -b_hP_h(x_i) - \ldots$$

$$- b_kP_k(x_i)\} \times 2\,\{-P_h(x_i)\}$$

luego

$$\sum_{i=1}^{n} y_iP_h(x_i) - b_h \sum_{i=1}^{n} P^2{}_h(x_i) = 0 \qquad \text{(los otros son ceros por ser ortogonales)}$$

de donde:

$$b_h = \frac{\sum_{i=1}^{n} y_iP_h(x_i)}{\sum_{i=1}^{n} P^2{}_h(x_i)}$$

Pero hay que conocer los polinomios. Vamos a ver los datos e incógnitas que tendremos:

En $P_0(x)$, sólo la incógnita C_{00}. En $P_1(x)$ dos: C_{10} y C_{11}. En $P_2(x)$ tres: C_{20}, C_{21}, C_{22} y así hasta P_k que tiene $C_{k1} \ldots C_{kk}$ o sea k+1 incógnitas. El nº total de coeficientes desconocidos será $\frac{(k+1)(k+2)}{2}$. Condiciones que tenemos: la de ortogonalidad, que son $\binom{k+1}{2} = \frac{(k+1)k}{2}$; luego sobran incógnitas, $\frac{(k+1)(k+2)}{2} - \frac{(k+1)k}{2} = k+1$. Vamos a fijar entonces, arbitrariamente, k+1 incógnitas y pondremos

$$C_{00} = C_{11} = C_{22} = \ldots = C_{kk} = 1$$

ahora el problema tiene solución pues hay el mismo número de incógnitas que de ecuaciones.

Vamos a calcular el polinómio $P_h(x)$. Pongamos las condiciones de ortogonalidad que cumple éste con los otros.

$$\sum_{i=1}^{n} P_0(x_i)\,P_h(x_i) = 0 \rightarrow \sum_{i=1}^{n} P_h(x_i) = 0 \qquad \text{(Por ser } P_0 = C_{00} = 1\text{)}$$

$$\sum_{i=1}^{n} P_1(x_i)\,P_h(x_i) = 0 \rightarrow \sum_{i=1}^{n} (c_{10} + x_i)P_h(x_i) = c_{10} \sum_{i=1}^{n} P_h(x_i) +$$

$$+ \sum_{i=1}^{n} x_iP_h(x_i) = \sum_{i=1}^{n} x_iP_h(x_i) = 0$$

$$\sum_{i=1}^{n} P_{h-1}(x_i)\,P_h(x_i) = 0 \rightarrow \sum_{i=1}^{n} x_i^{h-1}P_h(x_i) = 0$$

Tenemos h ecuaciones y las incógnitas de $P_h(x)$ son h, luego ya tenemos calculado el polinomio, y así todos los polinómics.

Vamos a llamar $a_r = \frac{1}{N}\sum_{i=1}^{n} x_i^{\,r}$ (momento de orden r respecto al origen) siendo N el n°de muestras.

$$\left.\begin{array}{ll} \text{Teníamos } \Sigma P_h(x_i) = 0\ , & \text{pondremos } \frac{1}{N}\,\Sigma\, P_h(x_i) = 0 \\ \text{Teníamos } \Sigma x_i P_h(x_i) = 0\ , & \text{pondremos } \frac{1}{N}\,\Sigma x_i P_h(x_i) = 0 \end{array}\right\}\ A$$

$$A = \left\{\begin{array}{l} \frac{1}{N}\ \Sigma(C_{ho}+C_{h_1}x_i+..+x_i^{h})=0 \\ \frac{1}{N}\ \Sigma(C_{ho}x_i+..+x_i^{h+1})\ =0 \\ \text{-----------------------} \\ \frac{1}{N}\ \Sigma(C_{ho}x_i^{h-1}+..+x_i^{2h-1})=0 \end{array}\right. = \left\{\begin{array}{l} C_{ho}+C_{h_1}a+..+a_h=0 \\ C_{ho}a_1+...+a_{h+1}=0 \\ \text{-----------------} \\ C_{ho}a_{h-1}+..+a_{2h-1}=0 \end{array}\right\}\ \begin{array}{l}\text{h ecuaciones -}\\ \text{con h incógnitas}\\ (C_{ij})\end{array}$$

Llamemos

$$\Delta(x) = \begin{vmatrix} 1 & a_1 & a_2 & \dots & a_h \\ a_1 & a_2 & a_3 & \dots & a_{h+1} \\ \text{--} & \text{--} & \text{--} & \text{--} & \text{--} \\ a_{h-1} & a_h & a_{h+1} & \dots & a_{2h-1} \\ 1 & x & x^2 & \dots & x^h \end{vmatrix}$$

y sea Δj el adjunto en $\Delta(x)$ de x^j y δj el menor de x^j

$$C_{hj} = \frac{(-1)\begin{vmatrix} 1 & a_1 & \dots & a_{h-1} \\ a_1 & a_2 & \dots & a_h \\ \text{--} & \text{--} & \text{--} & \text{--} \\ a_{h-1} & a_h & \dots & a_{2h-1} \end{vmatrix}}{\Delta h} = (-1)^{h-j+1}\,\frac{\delta j}{\Delta h}$$

$$C_{hj} = \frac{\Delta j}{\Delta h}$$

y queda

$$P_h(x) = \frac{\Delta o}{\Delta h} + \frac{\Delta 1}{\Delta h}\,x + \frac{\Delta 2}{\Delta h}\,x^2 + \,..\, + \frac{\Delta h-1}{\Delta h}\,x^{h-1} + x^h =$$

$$= \frac{1}{\Delta\, h}\left[\Delta_o + \Delta_1 x + \Delta_2 x^2 + \,..\, + \Delta_{h-1}x^{h-1} + \Delta h x^h\right] = \frac{\Delta(x)}{\Delta h}$$

4.2.5. Otras correlaciones

4.2.5.1. Correlación gráfico-analítica

Se trata de obtener una ley gráfica de correlación entre dos series de datos. Con este método, en vez de dibujar la nube de puntos que definen las parejas homólogas de las dos series, se define una cuadrícula en el plano xy, para lo cual se fijan los intervalos, tomándose como amplitud de los mismos, en cada uno de los ejes, un valor comprendido entre el 10% y el 20% del valor medio de cada serie. Como ya hemos dicho, en vez de dibujar los puntos definidos por las parejas de valores homólogos, se cuenta el número de puntos que tiene en cada una de las cuadrículas definidas.

Dando a cada cuadrícula un peso igual al número de puntos de cada una de ellas, se calculan los centros de gravedad por bandas horizontales y verticales, para el conjunto de la cuadrícula, que nos definen dos curvas para cada una de las bandas. Interpolando gráficamente estas dos curvas y, en caso necesario, utilizando también la correlación ortogonal, se obtiene la curva de correlación gráfica, que se utiliza principalmente para ajustar las partes extremas, máxima y mínima, de la recta de correlación ortogonal.

4.2.5.2. Método de las anomalías

Este método se utiliza para el trazado de isoyetas en zonas montañosas, haciendo la hipótesis de que los factores predominantes en la precipitación son la altitud y las características meteorológicas de las cuencas, por lo cual se subdivide la cuenca total en zonas parciales por criterios de afinidad hidrológica y correlaciones altitud - precipitación. En cada una de las zonas y en base anual o estacional, se establece la correlación altitud - precipitación, según puede verse en la figura I-4.21.

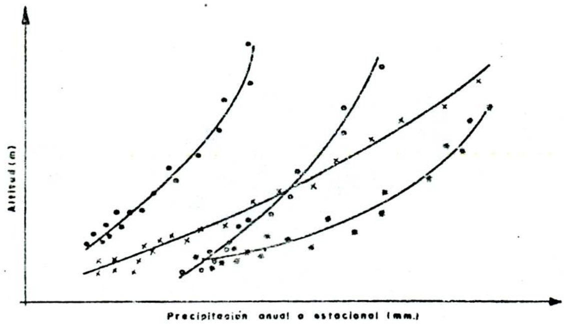


Fig.I-4.21

En cada una de las zonas se aplica a todas las estaciones la correlación precipitación - altitud, con lo cual se obtienen otras nuevas series, y las diferencias entre las series básicas y las obtenidas (anomalías) se pasan a un mapa topográfico y se trazan las isolíneas de las diferencias citadas. A partir de las correlaciones altitud - precipitación y el mapa de anomalías, se corrige o estima la precipitación en todos los puntos necesarios para el trazado de isoyetas.

4.3. DISTRIBUCION PROBABILISTICA DE LOS DATOS HIDROLOGICOS

4.3.1. Generalidades

4.3.1.1. Aplicación de la teoría de probabilidades a diversos fenómenos hidrológicos

Con las reservas que hemos indicado, la teoría de probabilidades tiene aplicaciones muy importantes para el estudio de los fenómenos hidrológicos. A continuación citamos diversos casos de aplicación.

Las curvas de caudales diarios clasificados o de duración de caudales, no es otra cosa que una función de distribución de caudales. Cuando el período a que se extienden las observaciones es reducido, habrá tramos de la curva en los que los caudales estarán fuertemente influenciados por los días anteriores, perdiendo, por consiguiente, el carácter puramente aleatorio.

Es interesante la determinación de las leyes de distribución de las aportaciones anuales para los estudios de regulación interanual, aunque ya hemos indicado que, frecuentemente, se observa cierta tendencia a agruparse los años secos en series, con mayor probabilidad que la que correspondería a la agrupación aleatoria de las aportaciones de varios años, obtenidas de la distribución de aportaciones anuales. Por esta razón, es útil investigar las leyes de distribución de las aportaciones correspondientes a agrupación de dos o más años sucesivos.

Para los estudios de explotación, se suelen investigar las leyes de distribución de las aportaciones en cada uno de los meses o estaciones del año. Si bien los resultados pueden -

ser un índice de las probabilidades de garantizar un cierto suministro durante esta época, pueden ocasionar fuertes errores, si se pretenden aplicar para estimar las probabilidades de un suministro próximo o para determinar las leyes de llenado o vaciado de un embalse, pues, en las aportaciones de cada mes, tienen sensible influencia las situaciones anteriores.

Por esta razón, para determinar las curvas de garantía de vaciado o de llenado de un embalse, es mucho más útil estudiar las leyes de distribución de las aportaciones acumuladas en 1, 2, 3, 4, ... meses a partir de cada mes del año, tanto hacia adelante como hacia atrás en el tiempo. Estos estudios serían muy laboriosos, si no se conocieran, con cierta aproximación, las fechas de comienzo y final del período de desembalse.

Es muy interesante la aplicación del cálculo de probabilidades al estudio de frecuencia de caudales de crecidas, aunque a veces los resultados de los estudios pueden resultar muy inciertos, por extrapolarse a períodos mucho más largos que aquél en el que se conocen datos. La aplicación de la estadística al conocimiento de los daños probables que los caudales de aguas altas pueden ocasionar en obras de corta duración (cimentación de obras, ataguías, etc.) suele exigir, a veces, el estudio de las leyes de distribución de caudales máximos solo durante ciertos períodos del año (verano). Como en los caudales de crecidas no muy distanciadas en el tiempo, pueden tener influencia las condiciones que han ocasionado otras crecidas anteriores, es frecuente estudiar la distribución, no de todos los caudales de crecidas, sino solamente la de los mayores caudales de cada año o mes (distribuciones de datos extremos). Otras veces, se consideran solamente los valores de caudales que sobrepasen determinado valor, ya que el exceso de datos pudiera falsear las leyes de distribución, que interesan fundamentalmente para los valores más elevados.

Para los caudales de estiaje puede ser interesante, en ciertos casos, no conocer la duración de caudales inferiores a valores determinados que se obtienen de las curvas de caudales clasificados, sino la frecuencia de estiajes acentuados, para lo que se estudia la ley de distribución de valores extremos de caudales de estiaje.

Las mismas leyes de distribución pueden aplicarse a los fenómenos meteorológicos: Precipitaciones anuales, precipitaciones máximas en determinado tiempo, distribuciones de temperaturas, etc., así como también a datos discontinuos, como días anuales de nevadas, días consecutivos sin lluvia, etc.

4.3.1.2. <u>Normas generales sobre los datos a emplear</u>

El conjunto de datos recopilados puede considerarse como una muestra del total de resultados posibles.

En la mayor parte de los casos, el número de datos es insuficiente para poder deducir la frecuencia de las situaciones poco probables, por demasiado altas o demasiado bajas, pero pueden dar una idea de los valores medios y de la dispersión de estos resultados.

Recordemos que si la población total está formada por un número muy elevado de sucesos, N, o bien infinito con una ley de frecuencias, f (ξ), para su definición son fundamentales los siguientes parámetros:

Media: $$\bar{\xi} = \int_{-\infty}^{\infty} \xi\, f(\xi)\, d\xi$$

ó $$\bar{\xi} = \sum_{1}^{N} \frac{\xi_i}{N}$$

Varianza:

$$\sigma^2 = \int_{-\infty}^{\infty} (\xi-\bar{\xi})^2\, f(\xi)\, d\xi$$

ó $$\sigma^2 = \sum_{1}^{N} \frac{(\xi_i-\bar{\xi})^2}{N} = \sum_{1}^{N} \frac{\xi_i^{\,2}}{N} - \bar{\xi}^2$$

Se llama desviación típica a la raíz cuadrada de la varianza:

$$\sigma = \sqrt{\sum_{1}^{N} \frac{(\xi_i-\bar{\xi})^2}{N}}$$

Así, pueden seguirse obteniendo los momentos de orden superior respecto a la media, de los que el tercero dá idea de la asimetría de la distribución:

$$\mu = \frac{1}{N} \sum_{1}^{N} (\xi_i-\bar{\xi})^3$$

Llamándose coeficiente de asimetría:

$$C_s = \frac{\mu}{\sigma^3}$$

Si se toman muestras, comprendiendo cada una n sucesos, la media de los valores de estos n sucesos

$$\bar{x} = \frac{1}{n} \sum_1^n x_i$$

tendrá una desviación standard respecto a la media general:

$$\sigma_n = \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{\sum_1^{\frac{N}{n}} (\bar{x}-\xi_i)^2}{\frac{N}{n}}}$$

Se demuestra que la desviación típica del conjunto N puede obtenerse, aproximadamente, de la correspondiente a la muestra de n sucesos, considerando en el denominador un elemento menos para compensar la desviación de las medias:

$$\sigma = \sqrt{\frac{\sum_1^n (x_i-\bar{x})^2}{n-1}}$$

y:

$$\mu = \frac{n}{(n-1)(n-2)} \sum_1^n (x_i-\bar{x})^3$$

Como veremos, estos valores son fundamentales para ajustar curvas de distribución analíticas.

Si el número de datos fuera suficiente, la ley de distribución de probabilidad podría representarse gráficamente, pero no convendría dar al valor más bajo una probabilidad $\frac{1}{n}$ ya que resultaría elevada en el último ($\frac{n}{n}=1$), pues puede haber valores más elevados que no hubieran aparecido en la muestra. Si se denomina m al puesto de cada suceso por orden creciente, parece criterio más riguroso dar a cada dato una probabilidad $\frac{2m-1}{2n}$, siendo también muy discutido este criterio, pues cada observación debería quedar colocada en la abcisa a la que correspondiera un 50% de probabilidad de rebasarse.

Con los datos así representados puede dibujarse, a sentimiento, una curva que pase, aproximadamente, por las inmediaciones de todos los puntos, o bien, ajustarse una curva analítica aproximada (figura I-4.22).

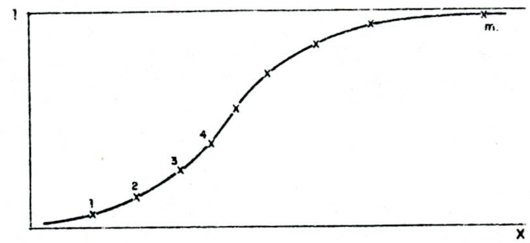


Fig.-I-4.22

Existen varios métodos para obtener la máxima aproximación entre los resultados obtenidos en una muestra y las funciones analíticas de distribución, como son, el método de igualar los momentos sucesivos de las leyes de frecuencia, el hacer mínimos los cuadrados de las diferencias o el método de la máxima verosimilitud, que resulta más complicado.

Las características más destacadas de las curvas de distribución y de su derivada, la curva de frecuencia, que rige en los fenómenos meteorológicos, son las siguientes: (fig. I-4.23)

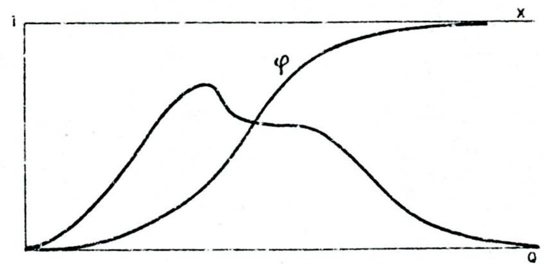


Fig.-I-4.23

Las curvas de frecuencia son forzosamente asimétricas, por no poder existir caudales negativos.

No es raro que en la curva de frecuencia se presenten dos máximos o, más generalmente, que sea poco apuntada, lo que equivale en la curva de distribución o probabilidad a una parte central casi recta. Lo mismo ocurre con otros fenómenos físicos, como en la resistencia a rotura de probetas de hormigón.

Los valores extremos conocidos (máximas crecidas, aportaciones de años secos, etc.), con frecuencia presentan una fuerte dispersión, saliéndose de la tendencia general de valores, - lo que hace difícil ajustar las curvas de distribución a sus zonas extremas, que a muchos efectos son las más importantes.

4.3.1.3. Verificación del carácter aleatorio de las series

Como hemos dicho anteriormente, si se toman muestras de n elementos x (por ejemplo caudales medios o máximos caudales), la desviación típica, σ_n, de las medias, $\overline{x}_m$, de cada muestra, - está ligada con la desviación típica de toda la población, σ, - por la expresión:

$$\sigma_n = \frac{\sigma}{\sqrt{n}}$$

De modo que, si se toman los caudales medios de cada n - años sucesivos y su desviación típica resulta sensiblemente superior a la que se deduciría de la expresión anterior, podrá -- afirmarse que hay dependencia entre las aportaciones de años sucesivos y que no será correcto aplicar el cálculo de probabilidades, - suponiendo que las aportaciones anuales se pueden agrupar aleatoriamente (fig. I-4.24.).

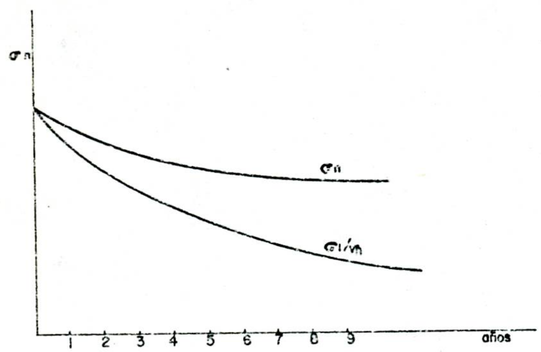


Fig.-I- 4.24

Como hemos dicho, la imprecisión de ajuste de curvas de probabilidad es muy grande, en caso de que se trate de extrapolar los resultados a períodos de recurrencia mucho más largos que el - correspondiente a la serie de datos, especialmente considerando que muchas veces los datos hidrográficos poco frecuentes se dispersan notablemente con respecto al conjunto de datos.

4.3.2. Bases teóricas

En muchos casos, los resultados de un fenómeno aleatorio pueden expresarse en forma numérica, como por ejemplo, la resistencia de unas probetas de hormigón, los pesos de personas, etc. En otros casos en que ésto no sea posible directamente podemos asignar valores numéricos convencionales a los posibles resultados, como por ejemplo, en un análisis de piezas de fundición, - designando por 1 el resultado "pieza buena" y por 0 el resultado "pieza mala". Supondremos pues en lo que sigue, que los resultados de un fenómeno o experimento aleatorio pueden expresarse por una cantidad numérica.

Se llama variable aleatoria a una cantidad numérica variable que expresa el resultado de un fenómeno aleatorio. La representaremos, en general, por una letra griega.

4.3.2.1. La función de distribución

4.3.2.1.1. Definición

Por $P(\xi = a)$ representaremos la probabilidad de que la variable aleatoria ξ tome el valor a.

Por $P(a < \xi \leqslant b)$ representaremos la probabilidad de que la variable tome un valor comprendido en el intervalo (a, b). - Si se conociera el valor de la probabilidad para todos los valores de a y b, es evidente que quedaría determinada la distribución de probabilidad de la variable ξ.

Sea x un número cualquiera y consideremos la probabilidad de que la variable ξ tome un valor menor o igual que x. Esta probabilidad será, pues, una función de x.

Llamaremos función de distribución de la variable aleatoria ξ, y la representaremos por F (x), a la probabilidad de - que la variable tome un valor menor o igual que x.

$$F(x) = P(\xi \leqslant x)$$

Esta función nos permite determinar la probabilidad de - que la variable tome un valor comprendido en un intervalo cualquiera (a , b). En efecto:

$$P(\xi \leq b) = P(\xi \leq a) + P(a < \xi \leq b)$$

y de aquí:

$$P(a < \xi \leq b) = F(b) - F(a)$$

4.3.2.1.2. Propiedades

a) Como una probabilidad es siempre un número no negativo, será:

$$F(b) - F(a) \geq 0, \text{ para } b \geq a.$$

luego la función F (x) es no decreciente

b) $$F(+\infty) = \lim_{x \to \infty} F(x) = 1$$

c) $$F(-\infty) = \lim_{x \to -\infty} F(x) = 0$$

4.3.2.2. Funciones de distribución discreta

Si la variable aleatoria solo puede tomar un número finito de valores $(x_1, x_2, \dots x_k)$ y las probabilidades de que se presenten cada uno de estos resultados son $(P_1, P_2, \dots P_k)$, la función de distribución será

$$F(x) = P(\xi \leq x) = \sum_{x_i \leq x} P_i$$

Llamaremos distribución discreta a una distribución de probabilidad de este tipo.

Una analogía mecánica muy útil consiste en suponer situados en los puntos $x_1, x_2, \dots x_k$ del eje real los pesos respectivos $P_1, P_2, \dots P_k$. (figura I-4.25).

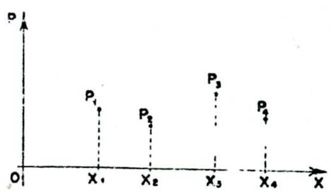


Fig. I- 4.25

El peso total del eje real será la unidad ya que

$$\sum_i P_i = 1$$

Podemos representar estos puntos en un sistema rectangular por sus coordenadas (x_i , P_i).

A la función discreta

$$f(x) = P(\xi = x)$$

la llamaremos función de densidad, y se verificará evidentemente

$$F(x) = \sum_{x_i \leq x} f(x_i)$$

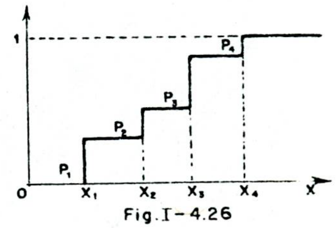


Fig. I-4.26

La función de distribución, F (x), vendrá entonces representada por una línea escalonada, que se incrementa en el escalón P_i en cada punto x_i (fig. I-4.26).

4.3.2.3. Funciones de distribución continua

Si la variable aleatoria puede tomar un número infinito de valores continuos en algún intervalo del eje real, sólo podrá hablarse de la probabilidad infinitesimal de que la variable tome un valor perteneciente a un intervalo infinitesimal, siendo 0 la probabilidad de que tome un valor determinado dentro de este intervalo.

Si recurrimos a la analogía mecánica introducida en la sección anterior, podemos suponer el eje real como una barra infinitamente delgada, de peso unidad, y con una densidad variable, f (x), en cada punto x.

La probabilidad de que la variable ξ tome un valor comprendido en el intervalo (x, x + dx) será entonces

$$P(x < \xi \leq x + dx) = f(x)\,dx,$$

y se verificará

$$F(x) = \int_{-\infty}^{x} f(x)\,dx$$

La función de densidad f (x) es, por consiguiente, la derivada de la función de distribución (figura I-4.27).

$$f(x) = F'(x)$$

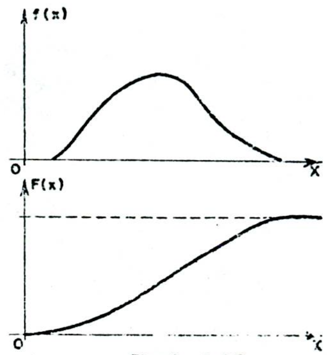


Fig. I-4.27

4.3.2.4. Parámetros característicos

4.3.2.4.1. Media o esperanza matemática

Llamaremos valor medio o esperanza matemática de la variable aleatoria, ξ, al centro de gravedad de la masa de la distribución, y lo representaremos por $E(\xi)$, o por la letra μ.

Para el caso discreto

$$E(\xi) = \sum_i P_i\, x_i$$

y para el continuo

$$E(\xi) = \int_{-\infty}^{\infty} x\, f(x)\,dx$$

Generalizando esta definición, llamaremos valor medio de una función, $\varphi(\xi)$, de una variable aleatoria, a los valores

$$E(\varphi(\xi)) = \sum_i P_i\, \varphi(x_i),$$

$$E(\varphi(\xi)) = \int_{-\infty}^{\infty} \varphi(x)\, f(x)\,dx,$$

para los casos discreto y continuo, respectivamente.

Para el caso de distribuciones en dos dimensiones definiremos igualmente:

$$E(\varphi(\xi,\eta)) = \sum_{i,j} P_{ij}\, \varphi(x_i, y_j)$$

$$E(\varphi(\xi,\eta)) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \varphi(x,y)\, f(x,y)\,dx\,dy.$$

Si $\varphi(\xi) = a\xi + b$, se obtiene fácilmente

$$E(a\xi + b) = a\,E(\xi) + b.$$

Si $\varphi(\xi,\eta) = \xi + \eta$, resulta:

$$E(\xi + \eta) = \sum_{i,j} P_{ij}(x_i + y_j) = \sum_i P_{ij} x_i + \sum_j P_{ij} y_j = E(\xi) + E(\eta),$$

$$E(\xi+\eta) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x+y) f(x,y)\,dx\,dy = \int_{-\infty}^{\infty} x f_1(x)\,dx + \int_{-\infty}^{\infty} y\, f_2(y)\,dy =$$

$$= E(\xi) + E(\eta).$$

Generalizando este resultado, se obtiene:

$$E(\xi_1 + \xi_2 + \ldots + \xi_n) = E(\xi_1) + E(\xi_2) + \ldots + E(\xi_n)$$

que es la regla de la suma para valores medios.

Si $\varphi(\xi,\eta) = \xi\eta$, será

$$E(\xi\eta) = \sum_{i,j} P_{ij}\, x_i x_j,$$

$$E(\xi\eta) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\, f(x,y)\,dx\,dy,$$

y si las variables son independientes, se verifica

$$P_{ij} = P_i\, P_j,$$

$$f(x,y) = f_1(x)\, f_2(y),$$

con lo que resulta

$$E(\xi\eta) = \sum_i P_i x_i \sum_j P_j y_j = \quad E(\xi) \cdot E(\eta),$$

$$E(\xi\eta) = \int_{-\infty}^{\infty} x f_1(x) dx \int_{-\infty}^{\infty} y\, f_2(y)\, dy = E(\xi) \cdot E(\eta).$$

que es la regla de multiplicación para variables independientes.

Generalizando, se obtiene:

$$E(\xi_1 \xi_2 \ldots \xi_n) = E(\xi_1)\ E(\xi_2) \ldots E(\xi_n)$$

siempre que $\xi_1\ \xi_2 \ldots \xi_n$ sean mutuamente independientes.

4.3.2.4.2. <u>Varianza</u>

Llamaremos varianza de la variable aleatoria, ξ, al momento de inercia de la masa de la distribución, y lo representaremos por $D^2(\xi)$.

Para el caso discreto

$$D^2(\xi) = E((\xi - \mu)^2) = \sum_i P_i (x_i - \mu)^2$$

y para el continuo

$$D^2(\xi) = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)\, dx$$

A la raíz cuadrada positiva de la varianza se la llama desviación típica o standard, y se representa por $D(\xi)$, o por la letra σ.

Una expresión útil para la varianza es

$$D^2(\xi) = E((\xi - \mu)^2) = E(\xi^2 - 2\mu\xi + \mu^2) =$$

$$= E(\xi^2) - 2\mu E(\xi) + \mu^2 = E(\xi^2) - \mu^2 =$$

$$= E(\xi^2) - E^2(\xi).$$

La varianza de una función lineal, $a\xi + b$, será:

$$D^2(a\xi + b) = E((a\xi + b - a\mu - b)^2) = E(a^2(\xi - \mu)^2) =$$

$$= a^2 E((\xi - \mu)^2) = a^2 D^2(\xi).$$

La varianza de una suma, $\xi + \eta$, de variables <u>independientes</u>, será:

$$D^2(\xi + \eta) = E((\xi + \eta)^2) - E^2(\xi + \eta) = E(\xi^2 + 2\xi\eta + \eta^2) -$$

$$- (E(\xi) + E(\eta))^2 = E(\xi^2) + E(\eta^2) + 2E(\xi\eta) -$$

$$- E^2(\xi) - E^2(\eta) - 2E(\xi)E(\eta)$$

y como hemos visto que en el caso de variables independientes

$$E(\xi\eta) = E(\xi)E(\eta)$$

resulta:

$$D^2(\xi + \eta) = E(\xi^2) - E^2(\xi) + E(\eta^2) - E^2(\eta) =$$

$$= D^2(\xi) + D^2(\eta)$$

Esta regla importante puede generalizarse al caso de un número arbitrario de variables independientes

$$D^2(\xi_1 + \xi_2 + \ldots + \xi_n) = D^2(\xi_1) + D^2(\xi_2) + \ldots + D^2(\xi_n)$$

4.3.2.4.3. <u>Moda</u>

Es el valor o valores de la variable de mayor probabilidad relativa o de mayor densidad de probabilidad.

En el caso de distribuciones discretas, la moda será:

$$\text{mod} = \sup_i\ p(x_i)$$

y en el caso continuo:

$$\text{mod} = \sup_x \; f(x).$$

En este último caso, supuestas las condiciones de continuidad, se obtendrá el valor resolviendo la ecuación

$$f'(x) = F''(x) = 0$$

4.3.2.4.4. Mediana

Es el valor de la variable que tiene una probabilidad $\frac{1}{2}$ de ser superado. En ambos casos se obtiene resolviendo la ecuación

$$F(x) = \frac{1}{2}$$

Como medidas de dispersión, además de la varianza o desviación típica, se empleó hace tiempo la desviación media, hoy en desuso:

$$\text{d.m.} = E(|\xi - \mu|)$$

4.3.2.4.5. Coeficientes de asimetría

Como medida de asimetría se emplea el coeficiente de asimetría relativa

$$\gamma_1 = \frac{\mu_3}{\sigma^3}$$

y también el coeficiente de asimetría absoluta

$$\gamma = \sqrt[3]{\mu_3}$$

En general, la sucesión de los momentos de una distribución o funciones de ellos, pueden ser empleados como parámetros característicos.

4.3.2.5. Momentos

4.3.2.5.1. Definición

Podemos definir, continuando con nuestra analogía mecánica, como momento de orden k, con relación al origen de coordenadas, el valor medio de ξ^k, que representaremos por α_k.

$$\alpha_k = E(\xi^k)$$

De forma análoga, podemos definir como momento de orden k, con relación a un punto de abscisa c, el valor medio de $(\xi - c)^k$.

En particular, se llama momento central de orden k, al valor medio de $(\xi - \mu)^k$, donde μ es la media de ξ, y se representará por μ_k.

$$\mu_k = E((\xi - \mu)^k)$$

4.3.2.5.2. Propiedades

Es fácil comprobar que se verifican las relaciones siguientes:

$$\mu_1 = 0$$

$$\mu_2 = \alpha_2 - \mu^2$$

$$\mu_3 = \alpha_3 - 3\mu\alpha_2 + 2\mu^3$$

$$\mu_4 = \alpha_4 - 4\mu\alpha_3 + 6\mu^2\alpha_2 - 3\mu^4$$

etc.,

y también

$$\alpha_1 = \mu$$

$$\alpha_2 = \mu_2 + \mu^2$$

$$\alpha_3 = \mu_3 + 3\mu\mu_2 + \mu^3$$

$$\alpha_4 = \mu_4 + 4\mu\mu_3 + 6\mu^2\mu_2 + \mu^4$$

etc.

4.3.2.5.3. Función generatriz

Definiremos como función generatriz de momentos de una variable aleatoria, ξ, a la función

$$\varphi(t) = E(e^{t\xi}),$$

donde t es una variable auxiliar imaginaria pura. El suponer que t sea imaginaria pura tiene como único objeto asegurar la convergencia de las expresiones que se utilizan después.

Para el caso de distribución discreta

$$\varphi(t) = \sum_i P_i\, e^{tx_i}$$

y para el de distribución continua

$$\varphi(t) = \int_{-\infty}^{\infty} e^{tx} f(x)\, dx$$

Si los momentos de orden k existen, obtenemos por derivación sucesiva en ambos casos

$$\varphi^{(k)}(0) = \alpha_k,$$

y el desarrollo en serie de Mac-Laurin de $\varphi(t)$ será:

$$\varphi(t) = \sum_{k=0}^{\infty} \frac{\alpha_k}{k!} t^k$$

luego el coeficiente de $\frac{t^k}{k!}$ nos dará el momento de orden k con relación al origen, por lo que $\varphi(t)$ recibe el nombre de función generatriz de momentos.

Si consideramos la función

$$\varphi_\mu(t) = E(e^{(\xi - \mu)t})$$

veremos que, en el correspondiente desarrollo, obtenemos los momentos centrales, por lo que se llama función generatriz de momentos centrales.

La relación entre las funciones φ y φ_μ viene dada por

$$\varphi_\mu(t) = E(e^{\xi t}\, e^{-\mu t}) = e^{-\mu t}\, E(e^{\xi t}) = e^{-\mu t}\, \varphi(t)$$

Si consideramos dos variables aleatorias independientes, ξ_1 y ξ_2, y llamamos respectivamente $\varphi_1(t)$, $\varphi_2(t)$ y $\varphi(t)$ a las funciones generatrices de las variables ξ_1, ξ_2 y de su suma $\xi_1 + \xi_2$, se verifica el siguiente teorema, fácilmente generalizable:

La función generatriz de la suma de dos variables independientes, es el producto de las funciones generatrices de estas dos variables.

En efecto:

$$\varphi(t) = E(e^{t(\xi_1+\xi_2)}) = E(e^{t\xi_1}\, e^{t\xi_2}) =$$

$$= E(e^{t\xi_1}) \,.\, E(e^{t\xi_2}) = \varphi_1(t)\, \varphi_2(t)$$

4.3.3. Funciones típicas de distribución utilizadas en hidrología

4.3.3.1. Introducción

Para facilitar los cálculos, es frecuente sustituir las curvas de distribución de caudales o aportaciones por funciones analíticas, debiendo advertirse que no existe, en general, fundamento teórico para que los fenómenos hidrológicos sigan tales leyes, que deben considerarse como empíricas.

Citamos a continuación los casos en que tienen aplicación más frecuente las leyes de distribución probabilística de caudales:

a) Hemos hablado de las curvas de duración de caudales, referentes a uno o varios años, que no es otra cosa que una ley de distribución de caudales, aunque, en el caso de un año, en los tramos extremos existirá una fuerte independencia entre los caudales. Suelen referirse las curvas de duración de caudales a los caudales medios diarios.

b) Caudales medios o aportaciones anuales, que pueden referirse a años oficiales o años hidráulicos, siendo frecuente, en las previsiones de explotación, estudiar la distribución de los caudales o aportaciones anuales, comenzando en cada mes.

c) Caudales medios o aportaciones mensuales, teniendo en cuenta la variación de los caudales a lo largo del año. Hay que considerar que las aportaciones mensuales, no solo dependen de la estación del año, sino de la situación en meses anteriores.

d) Aportaciones acumuladas a partir de una fecha determinada en 1, 2, 3, ... días, semanas o decenas, meses o años, distribución muy útil para estudios de explotación de embalses, tanto en épocas de desembalse como para estudiar las curvas de garantía de embalse.

e) Caudales mínimos, bien sean anuales o en determinados meses. Se emplean para el estudio de aprovechamientos sin regulación, en los que se exija gran garantía de suministro.

f) Caudales máximos, que pueden ser instantáneos o medios diarios. El estudio de esta distribución será necesario para dimensionar los aliviaderos de las presas, en cuyo caso conviene elegir los caudales máximos anuales, ya que, en otro caso, puede existir dependencia y prestar suma atención a los datos extremos. En caso de evaluación de los posibles daños de crecidas sobre obras provisionales (obras de desviación y ataguías), convendrá dar también importancia a los valores intermedios y estudiar la distribución de los caudales de crecidas por meses, considerando la posibilidad de concentrar los trabajos en los meses en que las crecidas sean menos probables.

Como escala para estudiar la distribución podrán medirse, los caudales medios en m^3/seg., las aportaciones en Hm^3. o m^3. por mes o año (al fin también son caudales), pudiendo emplearse caudales específicos (l/seg. y Km^2.) o aportaciones específicas (Hm^3/Km^2. y año) o bien, considerar las aportaciones representadas por la altura que supondría al repartirse por toda la cuenca, en mm., lo que tiene ventaja si han de ser comparadas con las precipitaciones.

4.3.3.2. Distribución de Gauss o normal

4.3.3.2.1. Distribución binomial

Consideremos un fenómeno o experimento aleatorio en el que sólo pueden presentarse los resultados A y B, con probabilidades respectivas p y q = 1 - p. Si realizamos <u>n</u> repeticiones independientes del experimento y llamamos <u>éxito</u> al suceso A, la probabilidad de obtener <u>x</u> éxitos será:

$$p_x = \binom{n}{x} p^x q^{n-x}$$

La distribución del número de éxitos tendrá, por consiguiente, como función de densidad

$$f(x) = \begin{cases} \binom{n}{x} p^x q^{n-x} & , \text{ para } x = 0, 1, 2 \ldots n \\ 0 & , \text{ para otros} \end{cases}$$

y como función de distribución:

$$F(x) = \sum_{i \leq x} \binom{n}{i} p^i q^{n-i} \qquad (i = 1, 2, \ldots n)$$

La función generatriz de momentos será

$$\varphi(t) = E(e^{\xi t}) = \sum_{x=0}^{n} \binom{n}{x} e^{xt} p^x q^{n-x} = (p e^t + q)^n$$

Los dos primeros momentos son:

$$\alpha_1 = n p$$

$$\alpha_2 = n (n - 1) p^2 + np$$

y por lo tanto, la esperanza y la varianza serán:

$$E(\xi) = n p$$

$$D^2(\xi) = npq$$

y los parámetros más usuales, media y desviación típica:

$$\mu = n\,p$$

$$\sigma = \sqrt{npq}$$

4.3.3.2.2. Distribución normal

Sea una variable aleatoria, ξ, que tiene una distribución binomial, y consideremos la nueva variable tipificada (1):

$$\lambda = \frac{\xi - np}{\sqrt{npq}}$$

La probabilidad de que la variable λ esté comprendida en el intervalo (a, b) será:

$$P\,(a<\lambda\leq b) = P\,(np + a\sqrt{npq}<\xi\leq np + b\sqrt{npq}) =$$

$$= \Sigma \binom{n}{i} p^i q^{n-i}$$

para todo valor de i tal que

$$np + a\sqrt{npq} < i \leqslant np + b\sqrt{npq}$$

De Moivre demostró en 1733 que, si p se mantiene constante mientras $n \to \infty$, la distribución discreta de λ se transforma en una distribución continua. Precisamente:

$$\lim_{n\to\infty} P\,(a < \lambda \leqslant b) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{z^2}{2}}\,dz.$$

Haciendo $a \to -\infty$, $b = x$, resulta:

$$\lim_{n\to\infty} P\,(\lambda \leqslant x) = \Phi\,(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{z^2}{2}}\,dz$$

(1) En general, se llama variable tipificada, a la obtenida de otra restándole la media y dividiendo por la desviación típica. La nueva variable tiene media cero y varianza unidad.

que es la llamada función de distribución normal o de Gauss. La función de densidad o de frecuencia será:

$$\Psi\,(x) = \Phi'\,(x)\ \frac{1}{\sqrt{2\pi}}\ e^{-\frac{x}{2}}$$

En la tabla adjunta se encuentran los valores de Φ (x) y Ψ (x). Para x < 0, pueden hallarse los valores correspondientes por medio de las relaciones: $\Phi\,(-x) = 1 - \Phi\,(x)$ y $\Psi\,(-x) = \Psi\,(x)$.

x	Φ (x)	Ψ (x)
0.0	0.50000	0.39894
0.1	0.53983	0.39695
0.2	0.57926	0.39104
0.3	0.61791	0.38139
0.4	0.65542	0.36827
0.5	0.69146	0.35207
0.6	0.72575	0.33322
0.7	0.75804	0.31225
0.8	0.78814	0.28969
0.9	0.81594	0.26609
1.0	0.84134	0.24197
1.1	0.86433	0.21785
1.2	0.88493	0.19419
1.3	0.90320	0.17137
1.4	0.91924	0.14973
1.5	0.93319	0.12952
1.6	0.94520	0.11092
1.7	0.95543	0.09405
1.8	0.96407	0.07895
1.9	0.97128	0.06562
2.0	0.97725	0.05399
2.1	0.98214	0.04398
2.2	0.98610	0.03547
2.3	0.98928	0.02833
2.4	0.99180	0.02239
2.5	0.99379	0.01753
2.6	0.99534	0.01358
2.7	0.99653	0.01042
2.8	0.99744	0.00792
2.9	0.99813	0.00595
3.0	0.99865	0.00443
3.1	0.99903	0.00327
3.2	0.99931	0.00238
3.3	0.99952	0.00172
3.4	0.99966	0.00123
3.5	0.99977	0.00087
3.6	0.99984	0.00061
3.7	0.99989	0.00042
3.8	0.99993	0.00029
3.9	0.99995	0.00020
4.0	0.99997	0.00013

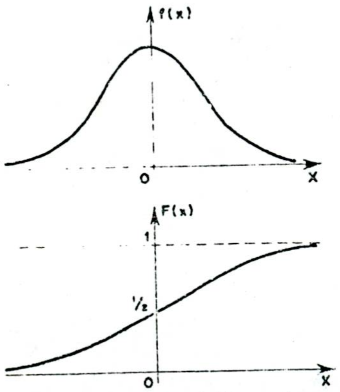


Fig. I- 4.28

La distribución es simétrica respecto del origen (figura I-4.28.) y, por consiguiente, todos los momentos impares son mulos. La media vale 0 y la varianza 1, en virtud del proceso de deducción, lo que puede comprobarse también por cálculo directo.

Consideremos la variable aleatoria

$$\xi = \sigma\lambda + \mu$$

donde $\sigma > 0$ y μ son constantes y λ tiene una distribución normal (seguiremos llamando normal a la distribución de ξ). La función de distribución de ξ será ahora:

$$\Gamma (x) = P (\xi \leq x) = P (\mu+\sigma\lambda \leq x) = P (\lambda \leq \frac{\xi-\mu}{\sigma}) =$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{x-\mu}{\sigma}} e^{-\frac{z^2}{2}} dz = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}\left(\frac{z-\mu}{\sigma}\right)^2} dz =$$

$$= \phi \left(\frac{x-\mu}{\sigma}\right)$$

y la función de frecuencia

$$f (x) = F' (x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} = \frac{1}{\sigma} \varphi \left(\frac{x-\mu}{\sigma}\right)$$

La función generatriz de momentos de la distribución normal generalizada será

$$\varphi (t) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tz} e^{-\frac{1}{2}\left(\frac{z-\mu}{\sigma}\right)^2} dz = e^{\mu t+\frac{1}{2}\sigma^2 t^2}$$

De aquí obtenemos

$$E (\xi) = \mu \qquad y \qquad D^2 (\xi) = \sigma^2$$

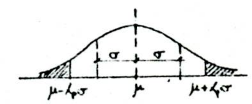
es decir, la media es μ y la desviación típica σ.

Una distribución normal queda, por consiguiente, definida si se conocen los dos parámetros, μ y σ. Diremos, para abreviar, que es N (μ, σ) y en el caso en que sea tipificada, N -- (0,1).

Todos los problemas numéricos planteados por una distribución N (μ, σ) pueden reducirse a los de la distribución tipificada, lo que reduce las tablas a las de la N (0,1).

Para un porcentaje p, el valor λ_p correspondiente de la distribución normal queda determinado por la condición

$$P (|\xi - \mu| > \lambda_p\sigma) = \frac{p}{100}$$



La probabilidad de que ξ se aparte de su media μ más de λ_p veces σ es p%.

λ_p en funcion de p		p en función de λ_p	
p	λ_p	λ_p	p
100	0.0000	0.0	100.000
95	0.0627	0.2	84.148
90	0.1257	0.4	68.915
85	0.1891	0.6	54.851
80	0.2533	0.8	42.371
75	0.3186	1.0	31.731
70	0.3853	1.2	23.014
65	0.4538	1.4	16.151
60	0.5244	1.6	10.960
55	0.5978	1.8	7.186
50	0.6745	2.0	4.550
45	0.7554	2.2	2.781
40	0.8416	2.4	1.640
35	0.9346	2.6	0.932
30	1.0364	2.8	0.511
25	1.1503	3.0	0.270
20	1.2816	3.2	0.137
15	1.4395	3.4	0.067
10	1.6449	3.6	0.032
5	1.9600	3.8	0.014
1	2.5758	4.0	0.006
0.1	3.2905		
0.01	3.8906		

Con la tabla anterior, de uso práctico, pueden hallarse los valores de las probabilidades de que una variable normal tome valores dentro o fuera de intervalos determinados.

Una propiedad muy importante de las distribuciones normales es la siguiente:

La suma de variables normales independientes es también normal.

En efecto: sean las variables ξ_1 y ξ_2 cuyas funciones generatrices son

$$\varphi_1 (t) = e^{\mu_1 t + \frac{1}{2}\sigma_1{}^2 t^2} \quad y \quad \varphi_2 (t) = e^{\mu_2 t + \frac{1}{2}\sigma_2{}^2 t^2}$$

La función generatriz de la suma será el producto de las funciones generatrices, luego

$$\varphi (t) = \varphi_1 (t) \, \varphi_2 (t) = e^{(\mu_1 + \mu_2) t + \frac{1}{2}(\sigma_1{}^2 + \sigma_2{}^2) t^2}$$

Vemos, por lo tanto, que $\varphi (t)$ corresponde a otra distribución normal de media

$$\mu = \mu_1 + \mu_2$$

y de varianza

$$\sigma^2 = \sigma_1{}^2 + \sigma_2{}^2$$

4.3.3.2.3. Ventajas e inconvenientes

Es la más utilizada en estadística.

La distribución normal tiene la ventaja de venir dada en función directa de los valores $\overline{x}$ y σ, características de la muestra de datos disponibles, así como también de encontrarse tabulada en casi todos los manuales.

Tiene los inconvenientes, por lo que se refiere a los fenómenos hidrológicos, de ser simétrica, dando valores negativos y frecuencias demasiado altas para valores muy altos o bajos, mientras la forma de la curva de frecuencia es muy apuntada y, en la realidad, las frecuencias alrededor de los valores medios suelen variar poco.

Para evitar la simetría de la distribución normal, se ha propuesto, por Galton, Hazen y otros, sustituir la variable aleatoria, z, por una función logarítmica

$$z = a \log (x - x_o) + b$$

calculándose los coeficientes a, b y x_o de manera que

$$\int_{-\infty}^{\infty} z \, f(z) \, dz = 0$$

$$\int_{-\infty}^{\infty} z^2 \, f(z) \, dz = 1$$

Pueden, también, compararse directamente $\overline{x}$ y σ, obtenidos directamente de la muestra, con los de la función normal transformada logarítmicamente, de donde tendremos:

$$\overline{x} = x_o + e^{A(\frac{A}{2} - b)}$$

$$\sigma^2 = -\overline{x}^2 + x_o{}^2 + 2x_o e^{A(\frac{A}{2} - b)} + e^{2A(A-b)}$$

$$\mu = (\overline{x} - x_o)^3 - 3(\overline{x} - x_o)\sigma^2 + e^{3A(\frac{3}{2}A - b)}$$

Donde $A = \dfrac{2,30259}{a}$

Deduciéndose:

$$\frac{\sigma^4}{\mu} = \frac{(\overline{x} - x_o)^3}{\sigma^2 + 3(\overline{x} - x_o)^2}$$

de donde se puede obtener x_o

$$A^2 = \log \left(1 + \frac{\sigma^2}{(\overline{x} - x_o)^2}\right)$$

$$b = \frac{1,1513}{a} - a \log (\overline{x} - x_o)$$

4.3.3.3. Distribución de Poisson

Consideremos de nuevo la distribución binomial

$$P_x = \binom{n}{x} p^x q^{n-x}$$

que podemos poner, mediante algunas transformaciones, en la forma

$$P_x = \frac{(np)^x}{x!} \left(1 - \frac{np}{n}\right)^n \frac{(1 - \frac{1}{n}) (1 - \frac{2}{n}) \ldots (1 - \frac{x-1}{n})}{(1 - p)^x}$$

Supongamos que la probabilidad p tiene un valor muy pe-- queño, tal que $p \to 0$ cuando $n \to \infty$ y de forma que

$$\lim_{\substack{n \to \infty \\ p \to 0}} np = \lambda$$

La probabilidad anterior, p_x, será, en el límite, igual a

$$\pi_x = \frac{\lambda^x}{x!} e^{-\lambda}$$

Para diversos valores de x y λ esta función está tabulada, figurando los valores hallados en la tabla adjunta.

λ \ x	0	1	2	3	4	5	6	7	8	9	10	11	12
0.1	.9048	.0905	.0045	.0002	.0000								
0.2	.8187	.1637	.0164	.0011	.0001	.0000							
0.3	.7408	.2222	.0333	.0033	.0002	.0000							
0,4	.6703	.2681	.0536	.0072	.0007	.0001	.0000						
0.5	.6065	.3033	.0758	.0126	.0016	.0002	.0000						
0.6	.5488	.3293	.0988	.0198	.0030	.0004	.0000						
0.7	.4966	.3476	.1217	.0284	.0050	.0007	.0001	.0000					
0.8	.4493	.3595	.1438	.0383	.0077	.0012	.0002	.0000					
0.9	.4066	.3659	.1647	.0494	.0111	.0020	.0003	.0000					
1.0	.3679	.3679	.1839	.0613	,0153	.0031	.0005	.0001	.0000				
1.1	.3329	.3662	.2014	.0738	.0203	.0045	.0008	.0001	.0000				
1.2	.3012	.3614	.2169	.0867	.0260	.0062	.0012	.0002	.0000				
1.3	.2725	.3543	.2303	.0998	.0324	.0084	.0018	.0003	.0001	.0000			
1.4	.2466	.3452	.2417	.1128	.0395	.0111	.0026	.0005	.0001	.0000			
1.5	.2231	.3347	.2510	.1255	.0471	.0141	.0035	.0008	.0001	.0000			
1.6	.2019	.3230	.2584	.1378	.0551	.0176	.0047	.0011	.0002	.0000			
1.7	.1827	.3106	.2640	.1496	.0636	.0216	.0061	.0015	.0003	.0001	.0000		
1.8	.1653	.2975	.2678	.1607	.0723	.0260	.0078	.0020	.0005	.0001	.0000		
1.9	.1496	.2842	.2700	.1710	.0812	.0309	.0098	.0027	.0006	.0001	.0000		
2.0	.1353	.2707	.2707	.1804	.0902	.0361	.0120	.0034	.0009	.0002	.0000		
2.2	.1108	.2438	.2681	.1966	.1082	.0476	.0174	.0055	.0015	.0004	.0001	.0000	
2.4	.0907	.2177	.2613	.2090	.1254	.0602	.0241	.0083	.0025	.0007	.0002	.0000	
2.6	.0743	.1931	.2510	.2176	.1414	.0735	.0319	.0118	.0038	.0011	.0003	.0001	.0000
2.8	.0608	.1703	.2384	.2225	.1557	.0872	.0407	.0163	.0057	.0018	.0005	.0001	.0000
3.0	.0498	.1494	.2240	.2240	.1680	.1008	.0504	.0216	.0081	.0027	.0008	.0002	.0001
3.2	.0408	.1304	.2087	.2226	.1781	.1140	.0608	.0278	.0111	.0040	.0013	.0004	.0001
3.4	.0334	.1135	.1929	.2186	.1858	.1264	.0716	.0348	.0148	.0056	.0019	.0006	.0002
3.6	.0273	.0984	.1771	.2125	.1912	.1377	.0826	.0425	.0191	.0076	.0028	.0009	.0003
3.8	.0224	.0850	.1615	.2046	.1944	.1477	.0936	.0508	.0241	.0102	.0039	.0013	.0004
4.0	.0183	.0733	.1465	.1954	.1954	.1563	.1042	.0595	.0298	.0132	.0053	.0019	.0006
5.0	.0067	.0337	.0842	.1404	.1755	.1755	.1462	.1044	.0653	.0363	.0181	.0082	.0034
6.0	.0025	.0149	.0446	.0892	.1339	.1606	.1606	.1377	.1033	.0688	.0413	.0225	.0113
7.0	.0009	.0064	.0223	.0521	.0912	.1277	.1490	.1490	.1304	.1014	.0710	.0452	.0264
8.0	.0003	.0027	.0107	.0286	.0573	.0916	.1221	.1396	.1396	.1241	.0993	.0722	.0481
9.0	.0001	.0011	.0050	.0150	.0337	.0607	.0911	.1171	.1318	.1318	.1186	.0970	.0728
10.0	.0000	.0005	.0023	.0076	.0189	.0378	.0631	.0901	.1126	.1251	.1251	.1137	.0948

λ \ x	13	14	15	16	17	18	19	20	21	22	23	24
5.0	.0013	.0005	.0002									
6.0	.0052	.0022	.0009	.0003	.0001							
7.0	.0142	.0071	.0033	.0014	.0006	.0002	.0001					
8.0	.0296	.0169	.0090	.0045	.0021	.0009	.0004	.0002	.0001			
9.0	.0504	.0324	.0194	.0109	.0058	.0029	.0014	.0006	.0003	.0001		
10.0	.0729	.0521	.0347	.0217	.0128	.0071	.0037	.0019	.0009	.0004	.0002	.0001

Hemos pasado, por consiguiente, de una distribución binomial a una distribución límite discreta, en la que la variable aleatoria puede tomar los valores 0, 1, 2, ..., x ... con probabilidades, p_x, que se llama distribución de Poisson.

La función generatriz será

$$\varphi(t) = \sum_0^\infty \pi_x e^{xt} = e^{-\lambda} \sum_0^\infty \frac{(\lambda e^t)^x}{x!} = e^{\lambda(e^t-1)}$$

Los dos primeros momentos son:

$$\alpha_1 = \lambda$$

$$\alpha_2 = \lambda + \lambda^2$$

y la esperanza y varianza

$$E(\xi) = \mu = \lambda$$

$$E^2(\xi) = \sigma^2 = \lambda$$

La suma de dos variables independientes, con distribuciones de Poisson de parámetros λ_1 y λ_2, tendrá por función generatriz de momentos

$$\varphi(t) = \varphi_1(t)\, \varphi_2(t) = e^{\lambda_1(e^t-1)}\, e^{\lambda_2(e^t-1)} = e^{(\lambda_1+\lambda_2)(e^t-1)}$$

que es la función generatriz de otra distribución de Poisson, - de parámetro $\lambda_1 + \lambda_2$.

En muchos casos prácticos, los sucesos se distribuyen -- aleatoriamente en el tiempo como por ejemplo los accidentes. Supongamos que la probabilidad de que un suceso ocurra en el in-- tervalo de tiempo (t, t+dt) es λdt.

Calculemos primero la probabilidad, $p_o(t)$, de que en el intervalo (o, t) no ocurra ningún suceso. Para ello consideraremos la probabilidad, $p_o(t+dt)$. Esta será igual a la probabilidad de que no ocurra ningún suceso en el intervalo (o, t), a la que hemos llamado p_o (t), ni en el intervalo (t, t+dt), que será la contraria de que ocurra algún suceso, es decir, $1 - \lambda dt$.

Tendremos entonces:

$$p_o\ (t+dt) = p_o\ (t)\ (1-\lambda dt)$$

y de aquí

$$p_o'\ (t) = -\lambda p_o\ (t)$$

Integrando esta ecuación resulta

$$p_o(t) = e^{-\lambda t}$$

habiendo determinado la constante de integración por la condición $p_o\ (0) = 1$.

Para hallar la probabilidad, p_x (t), de que ocurran x sucesos en el intervalo (0,t), procederemos de forma análoga, descomponiendo la probabilidad de que ocurran los x sucesos en el intervalo (0, t+dt) en una suma de dos probabilidades: la de que hayan ocurrido x sucesos en (0, t) y ninguno en (t, t+dt), y la de que hayan ocurrido x - 1 en (0, t) y uno en (t, t+dt). Por consiguiente:

$$p_x\ (t+dt) = p_x\ (t)(1-\lambda dt)+p_{x-1}(t)\ \lambda dt$$

y de aquí

$$p_x'\ (t) = -\lambda p_x\ (t) + \lambda p_{x-1}(t)$$

Llamando

$$p_x(t) = r_x\ (t)\ e^{-\lambda t} = p_o(t)\ r_x(t)$$

resulta

$$r_x'\ (t) = \lambda r_{x-1}(t)$$

y como $r_o\ (t) = 1$, por integraciones sucesivas hallamos r_1 (t), r_2 (t) ..., determinando las constantes de integración por la condición $r_x\ (0) = p_x\ (0) = 0$, para $x > 0$.

Así llegamos a la expresión general

$$p_x(t) = \frac{(\lambda t)^x}{x!}\ e^{-\lambda t}$$

El número de sucesos que pueden ocurrir en el intervalo de tiempo (o, t) tiene, por consiguiente, una distribución de Poisson con parámetro λt.

4.3.3.4. Distribución gamma o de Pearson

Recordemos primero que se define como función gamma y se representa por Γ (m) a la siguiente integral convergente:

$$\Gamma\ (m) = \int_0^\infty e^{-x}\ x^{m-1}\ dx$$

que goza de las propiedades

$$\Gamma\ (m) = (m-1)\ \Gamma\ (m-1)$$

$$\Gamma\ (1) = 1$$

$$\Gamma\ (\tfrac{1}{2}) = \sqrt{\pi}$$

Haciendo el cambio x = ay, resulta:

$$\int_0^\infty e^{-ay}\ y^{m-1}\ dy = \frac{\Gamma(m)}{a^m}$$

Llamaremos distribución gamma a la que tiene como función de densidad

$$f\ (x) = \frac{a^m}{\Gamma(m)}\ e^{-ax}\ x^{m-1}\ (0 < x < \infty)$$

y como función de distribución

$$F\ (x) = \frac{a^m}{\Gamma(m)} \int_0^\infty e^{-ax}\ x^{m-1}\ dx$$

La función generatriz de momentos será:

$$\varphi(t) = \frac{a^m}{\Gamma(m)} \int_0^\infty e^{-(a-t)x} x^{m-1} dx = \frac{a^m}{(a-t)^m} = \frac{1}{(1-\frac{t}{a})^m}$$

De aquí obtenemos los dos primeros momentos:

$$\alpha_1 = \frac{m}{a}$$

$$\alpha_2 = \frac{m(m+1)}{a^2}$$

La media y varianza serán, por consiguiente:

$$E(\xi) = \mu = \frac{m}{a}$$

$$D^2(\xi) = \sigma^2 = \frac{m}{a^2}$$

La suma de dos variables gamma independientes, de igual constante a, tendrá como función generatriz de momentos

$$\varphi(t) = \varphi_1(t)\,\varphi_2(t) = \frac{1}{(1-\frac{t}{a})^{m_1}} \frac{1}{(1-\frac{t}{a})^{m_2}} = \frac{1}{(1-\frac{t}{a})^{m_1+m_2}}$$

luego será otra variable gamma de igual constante a y parámetro $m_1 + m_2$

4.3.3.5. Distribución de Goodrich y Frechet

Considerando los valores de una serie de datos hidrológicos como una variable aleatoria, se ha comprobado que la distribución de estos valores puede ajustarse mediante una ley de Goodrich, cuya función de densidad de probabilidades es:

$$f(x) = \frac{1}{N} a (x-x_1)^{\frac{1}{n}-1} . e^{-a(x-x_1)^{\frac{1}{n}}}$$

siendo $x_1 \leq x < +\infty$, y n, x_1 y a, parámetros a determinar para cada muestra.

Igualando los tres primeros momentos de la muestra y de la distribución, queda:

$$\bar{x} = \frac{1}{a^n} \Gamma(n+1) + x_1 \quad \text{siendo } \Gamma(n+1) = \int_0^\infty e^{-x} . x^n dx$$

$$\sigma^2 = -(x_1 - \bar{x})^2 + \frac{1}{a^{2n}} \Gamma(2n+1)$$

$$\mu = (x_1 - \bar{x})^3 + \frac{1}{a^{3n}} \Gamma(3n+1) - 3\sigma^2 (x_1 - \bar{x}) ,$$

el cálculo de los coeficientes n, a y x_1, a partir de estas ecuaciones no resulta muy fácil, a veces, si no se tienen tablas o se recurre a procedimientos simplificados.

La forma de la función de distribución es:

$$F(x) = 1 - e^{-a(x-x_1)^{\frac{1}{n}}}$$

En algunos casos puede no ser conveniente la aplicación de la fórmula de Goodrich, especialmente si se estudian los mínimos valores probables, por presuponer un valor mínimo, x_1, y porque, aunque con poca probabilidad, pueden existir caudales inferiores a él.

La distribución de Frechet tiene la forma:

$$F(x) = e^{-(ax)^{-k}}$$

Es una expresión análoga a la de la ley de Goodrich, pero con signo negativo para el exponente k, por lo que, para calcular los coeficientes a y k se podría seguir un proceso análogo al indicado en líneas anteriores.

4.3.3.6. Distribución de Gumbel

Esta distribución de frecuencias se aplica, fundamentalmente, a valores extremos de datos hidrológicos. Su función de distribución tiene la forma siguiente:

$$F(x) = e^{-e^{-a(x-x_o)}}$$, siendo la función de densidad:

$$f(x) = ae^{-a(x-x_o)} \cdot e^{-e^{-a(x-x_o)}}$$, en donde, tanto

$\underline{a}$ como x_o, son parámetros a ajustar que se obtienen igualando los momentos de 1º y 2º orden, resultando las relaciones siguientes entre la media, $\bar{x}$, y la desviación típica:

$$\frac{1}{a} = 0{,}780\sigma$$

$$x_o = \bar{x} - \frac{0{,}577}{a}$$

Tiene la ley de Gumbel la ventaja respecto a la de Goodrich, de definirse sus coeficientes con facilidad, pero presenta el inconveniente de tener cierta rigidez al intervenir solamente dos parámetros.

Los valores negativos de la variable aleatoria, según este criterio de Gumbel, aún tendrían cierta probabilidad de suceder, lo que no es cierto para los caudales, que no pueden ser negativos. En cambio, en muchos casos, se aplica esta ley con cierta precisión para valores elevados de la variable aleatoria.

4.3.3.7. Distribuciones extremas

Frechet en 1927 (Tipo II) y Fisher-Tippett (Tipos I y III) en 1928 estudiaron, independientemente, la distribución de los valores extremos, encontrando que la distribución de los mayores valores de N (o de los N menores), cada uno de los cuales es escogido de uno de los "m" valores contenidos en cada N muestras, se aproxima a un límite (asintótico) de forma tal que "m" está siendo incrementado indefinidamente. El tipo de la forma del límite, depende del tipo de la distribución inicial de los valores Nm. Para tres diferentes tipos de distribución inicial, tres "distribuciones asintóticas extremas" pueden ser derivadas. Un estudio sistemático de las 3 asíntotas de los tipos correspondientes de la distribución inicial fué hecho por Von Mises.

4.3.3.7.1. Tipo I

Esta distribución resulta de cualquier distribución inicial del tipo exponencial, la cual converge a una función exponencial, como incremento de x. Ejemplos de tales distribuciones iniciales son la normal, qui-cuadrado y la lognormal. La densidad de probabilidad de la distribución del Tipo I es:

$$P(x) = \frac{1}{c} e^{-\frac{a+x}{c}} - e^{-\frac{a+x}{c}}$$

con $-\infty < x < +\infty$, donde x es la variación y $\underline{a}$ y $\underline{c}$ son los parámetros. La probabilidad acumulativa es:

$$P(X \leq x) = e^{-e^{-\frac{a+x}{c}}}$$

Por el método de los momentos, los parámetros fueron evaluados como

$$a = \gamma c - \mu$$

$$c = \frac{\sqrt{6}}{\pi}\sigma$$

donde $\gamma = 0{,}57721$ .... es la constante de Euler, μ es la media y σ es la desviación típica. La distribución tiene un coeficiente constante de asimetría, $C_s = 1{,}139$.

4.3.3.7.2. Tipo II

Esta distribución resulta de una distribución inicial del tipo Cauchy, en la cual no existen momentos de un cierto orden y más altos. La probabilidad acumulada es

$$P(X \leq x) = e^{-(\frac{\theta}{x})^k}$$

con $0 \leq x \leq \infty$, donde el parámetro θ es el mayor valor esperado, definido por la muestra del tamaño de $\underline{n}$ e incrementado con $\underline{n}$, y k es un orden de los momentos e independiente de $\underline{n}$.

4.3.3.7.3. Tipo III

Esta distribución resulta del tipo de la distribución -- inicial, en la cual x es limitado por $x \leq \varepsilon$.

La probabilidad acumulativa es:

$$P\ (X \leq x) = e^{-\left[(x-\varepsilon)/(\theta-\varepsilon)\right]^k}$$

con $-\infty < x \leq \varepsilon$

El parámetro k es el orden de la distribución más baja - de la función de probabilidad que no se anula para $x = \varepsilon$ y θ es el mayor valor esperado. En la práctica, la distribución del tipo I es, algunas veces, conocida como la distribución de Gumbel, desde que Gumbel aplicó por primera vez un análisis de frecuencia de crecidas. El tipo III es conocido como la distribución - de Weibull, que aplicó por primera vez esta distribución para - descripción de las resistencias a las fracturas de los materiales, aunque Gumbel también aplicó eso posteriormente para un -- análisis de frecuencia de estiajes.

4.3.3.8. Distribución de Galton

Se puede generalizar la ley de Gauss por cambios de va-- riables apropiadas. El más conocido de estos cambios de varia-- ble consiste en tomar como variable gaussiana el logaritmo, o -- una función lineal del logaritmo, de la variable estudiada. Se - obtiene así la ley de Galton o, por mejor decir, la ley de Gi-- brat-Gauss. Se la presenta tradicionalmente bajo la forma:

$$F\ (x) = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{z} e^{-z^2}\ dz$$

con $z = a \log (x-x_o) + b$

Es preferible adoptar una representación de la forma:

$$F\ (x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} e^{-\frac{u^2}{2}}\ du$$

con $u = a \log (x-x_o) + b$

Hemos introducido, para ciertas necesidades, un cambio - de variable en todo análogo pero con un parámetro menos. En esta ley, el logaritmo neperiano de la variable, (log x), sigue - una ley de Gauss. Se anota

$$F\ (x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y} e^{-\frac{1}{2}\left(\frac{y-\bar{y}}{\sigma_y}\right)^2} dy$$

con $y = \log x$

La función de densidad será

$$f\ (x) = \frac{1}{\sigma_y \sqrt{2\pi}}\ e^{-\frac{1}{2}\left(\frac{y-\mu_y}{\sigma_y}\right)^2}$$

y los diversos parámetros toman los valores

$$\mu = e^{\mu y + \frac{1}{2}\sigma_y^2}$$

$$\sigma = \mu\ (e^{\sigma_y^2} - 1)^{\frac{1}{2}}$$

$$M = e^{\mu y}\ (M = \text{mediana})$$

$$\alpha = (e^{3\sigma_y^2} - 3e^{\sigma_y^2} + 2)\ C_u^3$$

$$C_u = (e^{\sigma_y^2} - 1)^{\frac{1}{2}}$$

$$C_s = 3\ C_u + C_u^3$$

4.3.4. Recurrencia y probabilidad

El período de retorno, intervalo de recurrencia o repetición de un determinado acontecimiento, es el promedio de años - durante los cuales dicho acontecimiento será una vez igualado o sobrepasado.

Cabe hacer notar que los conceptos de período de retorno o de acontecimiento de un año tipo, no implican que un acontecimiento de una magnitud dada se produzca en intervalos constan--

tes ni aún aproximadamente constantes durante los N años. Ambos términos se refieren a una frecuencia media esperada durante un largo período de n años. Al estimar que un acontecimiento dado tendrá un período de retorno o recurrencia de 100 años, nada se dice sobre la cantidad en que este acontecimiento será sobrepasado. Es decir, el período de retorno (T_r) de un acontecimiento (X) fijado por el análisis de frecuencias, indica sólo el intervalo promedio entre acontecimientos mayores o iguales a X.

La probabilidad de que X se produzca en cualquier año, - está dada por la fórmula:

$$P = \frac{1}{T_r}$$

La probabilidad de que X se produzca por lo menos una -- vez, durante los n primeros años, viene dada por la expresión:

$$P_n = 1 - (1 - \frac{1}{T_r})^n$$

En la práctica, la cuestión se plantea, generalmente, del siguiente modo: sobre un elemento hidrológico cualquiera (altura de escala, precipitación, aportación específica), se recopilan, durante m años consecutivos, una serie de datos observados o determinados por un procedimiento indirecto aceptable y, a -- partir de estos datos, se determina para cada año un valor característico, importante bajo el punto de vista práctico. Por - ejemplo, la altura de agua máxima anual o la precipitación máxima, la altura de agua o la precipitación mínima y la altura de agua o la precipitación observada durante un número determinado de días al año. Estos valores anuales característicos, donde el número n es igual a los años de observación, varían según la sucesión de años secos o húmedos. El problema consiste ahora en - determinar la probabilidad de ocurrencia futura de un valor dado de esta característica o, inversamente, en determinar el valor correspondiente a una probabilidad dada.

Se definirá esta probabilidad diciendo que es igual a la frecuencia acumulada del valor correspondiente de la variable. Si se designa esta frecuencia por f se puede considerar que el valor en cuestión será ahora o después todos los $\frac{100}{f}$ años. Si -

por ejemplo, se lee sobre la curva de frecuencias acumuladas de altura de agua de una estación, que la frecuencia acumulada de altura de agua, h = 650 cm., es de 2%, este valor de 650 cm. será alcanzado o sobrepasado probablemente todos los $\frac{100}{2} = 50$ -- años. Mientras que $\frac{100}{f} < n$ este razonamiento es aceptable, pero si $\frac{100}{f} > n$ o, dicho de otra forma, el fenómeno se reproduce en unos períodos superiores a la duración de observación, la precisión no podrá ser tan buena como en el caso precedente.

La seguridad de los resultados aumenta, ciertamente, con la duración de las observaciones, es decir, con el número de éstas. Cada procedimiento permite prolongar la duración de observaciones en el pasado; por ejemplo, con la ayuda de fenómenos - que varían en el mismo sentido de los que se estudian, se podrá acrecentar la seguridad de las previsiones. Generalmente, una - tal extensión de la duración de observaciones no es posible y - entonces hemos de recurrir al razonamiento siguiente: la experiencia y la teoría muestran qu , si los valores son perfectamente aleatorios, es decir, cuando son independientes entre sí, la curva de repartición de frecuencias tiende hacia la curva de probabilidad de Gauss, cuando el número de observaciones aumenta. Para una serie de datos hidrológicos se aprecia, igualmente, que la curva tiende hacia una curva límite cuando el número de observaciones aumenta, pero no puede llegar a ser una curva de Gauss en razón del carácter asimétrico de la repartición de las observaciones (según las observaciones hechas hasta ahora se -- aproxima a una curva de Pearson de tercera categoría). Por consiguiente, si queremos calcular, a partir de una serie de observaciones, unos valores que se reproduzcan en períodos superiores a la duración de observación, se debe determinar la curva - de Pearson de tercer orden en la cual el coeficiente de variación, C_y, y el coeficiente de asimetría, C, corresponden a los valores calculados a partir de la serie de observaciones. Las - ordenadas de la curva acumulada, correspondientes a la curva de frecuencia así obtenida, nos darán los valores de la variable - correspondientes a unas probabilidades determinadas.

4.4. COMPROBACION DEL AJUSTE DE UNA DISTRIBUCION PROBABILISTICA

4.4.1. Test o contraste de hipótesis estadísticas

Hacer una hipótesis estadística sobre una población consiste en suponer conocido alguno o algunos de sus parámetros. - El aceptar o rechazar la hipótesis con cierta confianza medida numéricamente, en función de los valores de muestras observadas, se llama contraste de la hipótesis.

Supongamos que una variable estadística tiene una distribución aleatoria conocida. El método generalmente usado consiste en dividir el recorrido total de la variable en dos regiones, de aceptación y de rechace de la hipótesis, a cada una de las cuales corresponde una probabilidad. A la probabilidad correspondiente a la región de rechace se le llama nivel de significación. (figura I-4.29.).

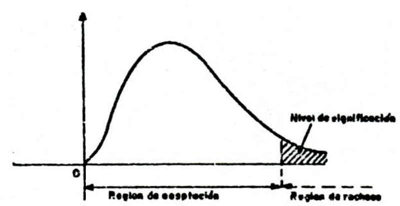


Fig I-4.29

Los niveles que suelen adoptarse en la práctica son los siguientes:

Casi significativo, cuando la probabilidad de rechazar la hipótesis es 5%.

<u>Significativo</u>, cuando la probabilidad es 1%.

<u>Muy significativo</u>, cuando la probabilidad es 1‰.

Esta forma de proceder se apoya en la admisión del principio de rechazar un resultado cuya probabilidad "a priori" era muy pequeña. Procediendo así, repetidamente, sólo nos equivocaremos al rechazar la hipótesis en un porcentaje de veces igual al nivel de significación. El aceptar la hipótesis, por otra parte, no quiere decir que ésta sea correcta, sino que no encontramos motivo suficiente para rechazarla.

Veremos a continuación algunos ejemplos de contraste de hipótesis estadísticas.

<u>Contraste de la media de una población normal</u>

Supongamos una población normal de la que conocemos σ. Vamos a contrastar la hipótesis de que la media es μ_0. Para ello extraemos una muestra $x_1, x_2, \ldots x_n$ y hallamos su media $\bar{x}$. Si la población origen tuviera, en efecto, μ_0 por media, la media muestral tendría una distribución

$$N\left(\mu_0 , \frac{\sigma}{\sqrt{n}}\right) .$$

Podemos elegir un intervalo centrado al que corresponda una probabilidad, por ejemplo del 95%, y considerar ésta como la región de aceptación (figura I-4.30). Si $\bar{x}$ pertenece a este intervalo, aceptaremos la hipótesis de que $\mu = \mu_0$, y en caso contrario la rechazaremos.

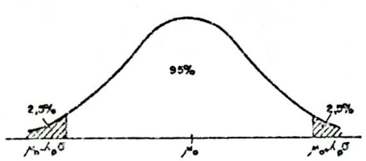


Fig I-4.30

Los valores correspondientes de λ_p se hallan en las tablas de la distribución normal y son:

Nivel casi significativo, 5% — $\lambda_p = 1,96$
Nivel significativo, 1% — $\lambda_p = 2,58$
Nivel muy significativo, 1‰ — $\lambda_p = 3,29$

Si $\bar{x}$ no perteneciera, por ejemplo, al intervalo (figura I-4.30)

$$(\mu_0 - 1,96\sigma , \quad \mu_0 + 1,96\sigma)$$

se dirá que la desviación de $\bar{x}$ sobre μ_0 es casi significativa.

<u>Contraste de la varianza de una población normal</u>

Supongamos una población normal en la que queremos contrastar la hipótesis de que su desviación típica es σ_0. Para ello extraemos una muestra $x_1, x_2, \ldots x_n$ y hallamos su varianza s^2. Sabemos que ns^2/σ^2 tiene una distribución χ^2 con n - 1 grados de libertad. Elegido el nivel de significación, se hallan en las tablas de χ^2 los valores correspondientes de χ^2_p. (figura I-4.31).

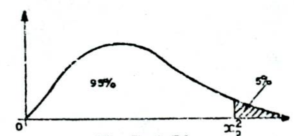


Fig I-4.31

<u>Contraste de la diferencia entre medias de una población normal</u>

Se presenta muchas veces el problema de contrastar la hipótesis de que dos muestras proceden de la misma población. Sean dos muestras de extensiones n_1 y n_2 y tratamos de contrastar la hipótesis de que la diferencia entre sus medias, $\bar{x}$ e $\bar{y}$, es debida al azar.

Si conocemos la desviación típica, σ, la distribución de la variable aleatoria, $\bar{x} - \bar{y}$, será

$$N\left(0, \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)$$

por lo que podemos hallar fácilmente la región de aceptación mediante la distribución normal.

Si no se conoce la desviación típica, sabemos que $n_1 s_1^2/\sigma^2$ tiene una distribución χ^2 con n_1 - 1 grados de libertad. La distribución de la variable

$$\frac{n_1 s_1^2 + n_2 s_2^2}{\sigma^2}$$

será una χ^2 con $n_1 + n_2 - 2$ grados de libertad, y, finalmente, la de la variable

$$t = \frac{\bar{x} - \bar{y}}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} : \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{\sigma^2 (n_1 + n_2 - 2)}} = (\bar{x} - \bar{y}) \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{(n_1 + n_2)(n_1 s_1^2 + n_2 s_2^2)}}$$

será una t de Student con $n_1 + n_2 - 2$ grados de libertad, con lo que podemos hallar la región de aceptación de la hipótesis.

Contraste de la incorrelación entre dos poblaciones normales

Otro problema que se presenta en muchos casos es el de determinar si entre dos poblaciones normales no existe correlación, es decir, si su coeficiente de correlación es 0. Se demuestra entonces que la variable

$$t = \sqrt{n-2} \; \frac{r}{1 - r^2}$$

donde r es el coeficiente de correlación de la muestra y n su extensión, tiene una distribución de Student con $n - 2$ grados de libertad, lo que nos permite determinar la región de aceptación para r.

Contraste de muestras de atributos

Consideremos una población en la que existe una probabilidad constante, p, de que cada uno de sus individuos tenga un determinado atributo.

Tendremos, en este caso, una distribución binomial para el número de éxitos, que para $n > 30$ puede asimilarse a una distribución normal. El número esperado de éxitos será np y su desviación típica $\sqrt{npq}$. Los intervalos centrados de aceptación serán por lo tanto:

95% : $np - 1{,}96\sqrt{npq}$, $np + 1{,}96\sqrt{npq}$

99% : $np - 2{,}58\sqrt{npq}$, $np + 2{,}58\sqrt{npq}$

99,9% : $np - 3{,}29\sqrt{npq}$, $np + 3{,}29\sqrt{npq}$

4.4.2. Test de χ^2 de Pearson

4.4.2.1. Distribución χ^2

Consideremos una variable, ξ, de distribución N (0,1), y hallemos la distribución de la variable $\eta = \xi^2$. Su función de distribución será:

$$G(x) = P(\eta \leq x) = P(\xi^2 \leq x) = P(-\sqrt{x} < \xi \leq \sqrt{x}) = \frac{1}{\sqrt{2\pi}} \int_{-\sqrt{x}}^{\sqrt{x}} e^{-\frac{1}{2} y^2} dy$$

y su función de densidad

$$g(x) = G'(x) = \frac{1}{\sqrt{2\pi}} \left(\frac{1}{2\sqrt{x}} e^{-\frac{x}{2}} + \frac{1}{2\sqrt{x}} e^{-\frac{x}{2}} \right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x}{2}} x^{\frac{1}{2}-1}$$

que es la función de densidad de una distribución gamma para $a = \frac{1}{2}$, $m = \frac{1}{2}$.

Su función generatriz de momentos será

$$\varphi(t) = \frac{1}{(1-2t)^{1/2}}$$

Sean ahora las n variables independientes $\xi_1, \xi_2, \ldots \xi_n$, todas N (0,1).

La nueva variable

$$\chi^2 = \xi_1^2 + \xi_2^2 + \ldots + \xi_n^2$$

tendrá como función generatriz de momentos

$$\varphi(t) = \frac{1}{(1-2t)^{1/2}} \frac{1}{(1-2t)^{1/2}} \cdots \frac{1}{(1-2t)^{1/2}} = \frac{1}{(1-2t)^{n/2}}$$

luego χ^2 tendrá una distribución gamma para $a = \frac{1}{2}$, $m = \frac{n}{2}$. Esta distribución se llama χ^2 de Pearson y tiene como funciones de densidad y distribución

$$k_n(\chi^2) = \frac{1}{2^{n/2}\,\Gamma(\frac{n}{2})} e^{-\frac{\chi^2}{2}} (\chi^2)^{\frac{n}{2}-1}$$

$$K_n(\chi^2) = \frac{1}{2^{n/2}\,\Gamma(\frac{n}{2})} \int_0^{\chi^2} e^{-\frac{x^2}{2}}\, x^{n-2}\, dx$$

Al parámetro n se le suele llamar "grados de libertad".

Para un porcentaje p, el valor $\chi^2{}_p$ de la distribución χ^2 se define por la condición

$$P(\chi^2 > \chi^2{}_p) = \frac{P}{100}$$

Por consiguiente, la probabilidad de que χ^2 supere a $\chi^2{}_p$ es igual a p%. En la tabla adjunta se dan los valores de $\chi^2{}_p$, en función de n y p

n	$\chi^2{}_p$ en función de n y p													
	p=99	98	95	90	80	70	50	30	20	10	5	2	1	0.1
1	0.000	0.001	0.004	0.016	0.064	0.148	0.455	1.074	1.642	2.706	3.841	5.412	6.635	10.827
2	0.020	0.040	0.103	0.211	0.446	0.713	1.386	2.408	3.219	4.605	5.991	7.824	9.210	13.815
3	0.115	0.185	0.352	0.584	1.005	1.424	2.366	3.665	4.642	6.251	7.815	9.837	11.341	16.268
4	0.297	0.429	0.711	1.064	1.649	2.195	3.357	4.878	5.989	7.779	9.488	11.668	13.277	18.465
5	0.554	0.752	1.145	1.610	2.343	3.000	4.351	6.064	7.289	9.236	11.070	13.388	15.086	20.517
6	0.872	1.134	1.635	2.204	3.070	3.828	5.348	7.231	8.558	10.645	12.592	15.033	16.812	22.457
7	1.239	1.564	2.167	2.833	3.822	4.671	6.346	8.383	9.803	12.017	14.067	16.622	18.475	24.322
8	1.646	2.032	2.733	3.490	4.594	5.527	7.344	9.524	11.030	13.362	15.507	18.168	20.090	26.125
9	2.088	2.532	3.325	4.168	5.380	6.393	8.343	10.656	12.242	14.684	16.919	19.679	21.666	27.877
10	2.558	3.059	3.940	4.865	6.179	7.267	9.342	11.781	13.442	15.987	18.307	21.161	23.209	29.588
11	3.053	3.609	4.575	5.578	6.989	8.148	10.341	12.899	14.631	17.275	19.675	22.618	24.725	31.264
12	3.571	4.178	5.226	6.304	7.807	9.034	11.340	14.011	15.812	18.549	21.026	24.054	26.217	32.909
13	4.107	4.765	5.892	7.042	8.634	9.926	12.340	15.119	16.965	19.312	22.362	25.472	27.689	34.526
14	4.660	5.366	6.571	7.790	9.467	10.821	13.339	16.222	18.151	21.064	23.685	26.873	29.141	36.123
15	5.229	5.985	7.261	8.547	10.307	11.721	14.339	17.322	19.311	22.307	24.996	28.259	30.578	37.697
16	5.612	6.614	7.962	9.312	11.152	12.624	15.338	18.418	20.465	23.542	26.296	29.633	32.000	39.252
17	6.408	7.255	8.672	10.085	12.002	13.531	16.338	19.511	21.615	24.769	27.587	30.995	33.409	40.790
18	7.015	7.906	9.390	10.875	12.857	14.440	17.338	20.601	22.760	25.989	28.869	32.346	34.805	42.312
19	7.633	8.567	10.117	11.651	13.716	15.352	18.338	21.689	23.900	27.204	30.144	33.687	36.191	43.820
20	8.260	9.237	10.851	12.443	14.578	16.266	19.337	22.775	25.038	28.412	31.410	35.020	37.566	45.315
21	8.897	9.915	11.591	13.240	15.445	17.182	20.337	23.858	26.171	29.615	32.671	36.343	38.932	46.797
22	9.542	10.600	12.338	14.041	16.314	18.101	21.337	24.939	27.301	30.813	33.924	37.659	40.289	48.268
23	10.196	11.293	13.091	14.848	17.187	19.021	22.337	26.018	28.429	32.007	35.172	38.968	41.638	49.728
24	10.856	11.992	13.848	15.659	18.062	19.943	23.337	27.096	29.553	33.196	36.415	40.270	42.980	51.179
25	11.524	12.697	14.611	16.473	18.940	20.867	24.337	28.172	30.675	34.382	37.652	41.566	44.314	52.620
26	12.198	13.409	15.379	17.292	19.820	21.792	25.336	29.246	31.795	35.563	38.885	42.856	45.642	54.052
27	12.879	14.125	16.151	18.114	20.703	22.719	26.336	30.319	32.912	36.741	40.113	44.140	46.963	55.476
28	13.565	14.847	16.9[illegible]	18.939	21.588	23.647	27.336	31.391	34.027	37.916	41.337	45.419	48.278	56.893
29	14.256	15.574	17.708	19.768	22.475	24.577	28.336	32.461	35.139	39.087	42.557	46.693	49.588	58.302
30	14.953	16.306	18.493	20.599	23.364	25.508	29.336	33.530	36.250	40.256	43.773	47.962	50.892	59.703

Para n > 30, el valor de $\chi^2{}_p$ se calcula con las tablas de la distribución normal, teniendo en cuenta la relación

$$P(\chi^2 > \chi^2{}_p) = 1 - \phi\left(\frac{\chi^2{}_p - n}{\sqrt{2n}}\right)$$

4.4.2.2. Estudio del test χ^2

Sea una población de la que hemos extraído la muestra x_1, x_2, ... x_n y queremos contrastar la hipótesis de que la población tiene una distribución F (x).

Si dividimos el recorrido total de la variable en un número finito de partes r, y llamamos P_i a la probabilidad de que la variable tome un valor comprendido en el intervalo i, y f_i al número de observaciones comprendidas en dicho intervalo, tendremos:

$$p_1 + p_2 + \dots + p_r = 1$$

$$f_1 + f_2 + \dots + f_r = n$$

Cada f_i representa la frecuencia observada en n repeticiones del suceso de probabilidad, p_i, por lo que f_i tendrá una distribución binomial de media, n p_i, y desviación típica, $\sqrt{np_iq_i} = \sqrt{np_i(1-p_i)} \approx \sqrt{np_i}$, supuesto que p_i es pequeño. Las variables

$$z_i = \frac{f_i - np_i}{\sqrt{np_i}}$$

tendrán una distribución asintónicamente N (0, 1), y la suma

$$\chi^2 = z_1{}^2 + z_2{}^2 + \dots + z_r{}^2$$

tendrá una distribución χ^2 con r -1 grados de libertad, puesto que entre las variables z_i existe la relación

$$\sum_i^r P_i\, z_i = 0$$

Puede, por consiguiente, hallarse la región de rechace para un determinado nivel de significación.

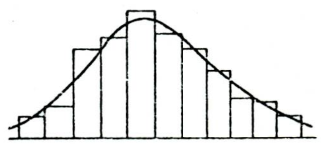


Fig I-4.32

En las aplicaciones prácticas conviene que $np_i \geq 10$, lo que puede conseguirse escogiendo los intervalos de forma que cada uno contenga, al menos, 10 valores observados (figura I-4.32.). Cuando las ob servaciones son tan pocas que no puede hacerse ésto, los resultados obtenidos mediante este test deben considerarse con precaución.

Si la distribución dependiese de varios parámetros desconocidos que se han estimado a partir de la muestra, como veremos más adelante, se demuestra que la expresión

$$\sum_{i}^{r} \left(\frac{f_i - np_i}{\sqrt{np_i}} \right)^2$$

tiene una distribución χ^2 con $r - k - 1$ grados de libertad, donde k es el número de parámetros que se han estimado.

4.4.3. Test de Kolmogoroff

4.4.3.1. Distribución de Kolmogoroff

Sea $F(x)$ la función de distribución de una población, ξ, y sea $F_n(x)$ la distribución estadística de n valores observados. Llamemos

$$D_n = \sup |F(x) - F_n(x)|$$

y consideremos la función

$$\Phi_n(z) = P(\sqrt{n}\ D_n \leq z)$$

Kolmogoroff ha demostrado que

$$\lim_{n \to \infty} \Phi_n(z) = K(z) = \sum_{-\infty}^{\infty} (-1)^k\ e^{-2k^2 z^2}$$

que es una distribución asintótica tabulada, cuyos valores se reflejan en la tabla siguiente.

z	K(z)	z	K(z)	z	K(z)	z	K(z)	z	k(z)
0,32	.0000	0,72	.3223	1,12	.8374	1,52	.9803	1,92	.8897
0,33	.0001	0,73	.3391	1,13	.8445	1,53	.9815	1,93	.9988
0,34	.0002	0,74	.3560	1,14	.8514	1,54	.9826	1,94	.9989
0,35	.0003	0,75	.3728	1,15	.8580	1,55	.9836	1,95	.9990
0,36	.0005	0,76	.3896	1,16	.8644	1,56	.9846	1,96	.9991
0,37	.0008	0,77	.4064	1,17	.8706	1,57	.9855	1,97	.9991
0,38	.0013	0,78	.4230	1,18	.8765	1,58	.9864	1,98	.9992
0,39	.0019	0,79	.4395	1,19	.8823	1,59	.9873	1,99	.9993
0,40	.0028	0,80	.4559	1,20	.8877	1,60	.9880	2,00	.9993
0,41	.0040	0,81	.4720	1,21	.8930	1,61	.9888	2,01	.9994
0,42	.0055	0,82	.4880	1,22	.8981	1,62	.9895	2,02	.9994
0,43	.0074	0,83	.5038	1,23	.9030	1,63	.9902	2.03	.9995
0,44	.0097	0,84	.5194	1,24	.9076	1,64	.9908	2,04	.9995
0,45	.0126	0,85	.5347	1,25	.9121	1,65	.9914	2,05	.9996
0,46	.0160	0,86	.5497	1,26	.9164	1,66	.9919	2,06	.9996
0,47	.0200	0,87	.5645	1,27	.9205	1,67	.9924	2,07	.9996
0,48	.0247	0,88	.5791	1,28	.9245	1,68	.9929	2,08	.9996
0,49	.0300	0,89	.5933	1,29	.9283	1,69	.9934	2,09	.9997
0,50	.0361	0,90	.6073	1,30	.9319	1,70	.9938	2,10	.9997
0,51	.0428	0,91	.6209	1,31	.9354	1,71	.9942	2,11	.9997
0,52	.0503	0,92	.6343	1,32	.9387	1,72	.9946	2,12	.9997
0,53	.0585	0,93	.6473	1,33	.9418	1,73	.9950	2,13	.9998
0,54	.0675	0,94	.6601	1,34	.9449	1,74	.9953	2,14	.9998
0,55	.0772	0,95	.6725	1,35	.9478	1,75	.9956	2,15	.9998
0,56	.0876	0,96	.6846	1,36	.9505	1,76	.9959	2,16	.9998
0,57	.0987	0,97	.6964	1,37	.9531	1,77	.9962	2,17	.9998
0,58	.1104	0,98	.7079	1,38	.9556	1,78	.9965	2,18	.9999
0,59	.1228	0,99	.7191	1,39	.9580	1,79	.9967	2,19	.9999
0,60	.1357	1,00	.7300	1,40	.9603	1,60	.9969	2,20	.9999
0,61	.1492	1,01	.7406	1,41	.9625	1,81	.9971	2,21	.9999
0,62	.1662	1,02	.7508	1,42	.9646	1,82	.9973	2,22	.9999
0,63	.1778	1,03	.7608	1,43	.9665	1,83	.9975	2,23	.9993
0,64	.1927	1,04	.7704	1,44	.9684	1,84	.9977	2,24	.9999
0,65	.2080	1,05	.7798	1,45	.9702	1,85	.9979	2,25	.9999
0,66	.2236	1,06	.7889	1,46	.9718	1,85	.9980	2,26	.9999
0,67	.2396	1,07	.7976	1,47	.9734	1,87	.9981	2,27	.9999
0,68	.2558	1,08	.8061	1,48	.9750	1,88	.9983	2,28	.9999
0,69	.2722	1,09	.8143	1,49	.9764	1,89	.9984	2,29	.9999
0,79	.2888	1,10	.8223	1,50	.9778	1,90	.9985	2,30	.9999
0,71	.3055	1,11	.8299	1,51	.9791	1,91	.9986	2,31	1.0000

4.4.3.2. Estudio del test de Kolmogoroff

Para contrastar la bondad de un ajuste puede utilizarse el test de Kolmogoroff. Si $F(x)$ es la función de distribución de una población, y $F_n(x)$ la distribución estadística de n valores observados, el valor $1-k(z)$ nos da la idea de la bondad de la distribución $F_n(x)$.

Los valores correspondientes a los límites de aceptación para $\sqrt{n}\, D_n$ son:

95% — $z_1 = 1,36$

99% — $z_2 = 1,63$

99,9% — $z_3 = 1,95$

Si $D_n < z/\sqrt{n}$ se acepta la hipótesis, rechazándola en caso contrario.

4.4.4. Test de Student

4.4.4.1. Distribución, t, de Student

Sean ahora dos variables independientes, ξ y η, la primera con una distribución $N(0,1)$ y la segunda con una distribución χ^2 con n grados de libertad. Se demuestra, mediante algunas transformaciones, que la variable

$$\theta = \sqrt{n}\ \frac{\xi}{\sqrt{\eta}}$$

tiene como función de densidad

$$S_n(t) = \frac{1}{\sqrt{n\pi}}\ \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}\ \frac{1}{\left(1+\frac{t^2}{n}\right)^{\frac{n+1}{2}}}$$

Esta distribución se llama t de Student.

Para un porcentaje, p, el valor t_p de la distribución t, se define por la condición.

$$P(|t| > t_p) = \frac{P}{100}$$

Por consiguiente, la probabilidad de que t difiera de su media cero, en ambos sentidos, en más de t_p es igual a p%.

Estos valores se tabulan en el cuadro siguiente.

n	t_p en función de n y p												
	p=90	80	70	60	50	40	30	20	10	5	2	1	0.1
1	0.158	0.325	0.510	0.727	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657	636.619
2	0.142	0.289	0.445	0.617	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	31.598
3	0.137	0.277	0.424	0.584	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	12.941
4	0.134	0.271	0.414	0.569	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	8.610
5	0.132	0.267	0.408	0.559	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	6.859
6	0.131	0.265	0.404	0.553	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.959
7	0.130	0.263	0.402	0.549	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	5.405
8	0.130	0.262	0.399	0.546	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	5.041
9	0.129	0.261	0.398	0.543	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.781
10	0.129	0.260	0.397	0.542	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.587
11	0.129	0.260	0.396	0.540	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.437
12	0.128	0.259	0.395	0.539	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	4.318
13	0.128	0.259	0.394	0.538	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	4.221
14	0.128	0.258	0.393	0.537	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	4.140
15	0.128	0.258	0.393	0.536	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	4.073
16	0.128	0.258	0.392	0.535	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	4.015
17	0.128	0.257	0.392	0.534	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.965
18	0.127	0.257	0.392	0.534	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.922
19	0.127	0.257	0.391	0.533	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.883
20	0.127	0.257	0.391	0.533	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.850
21	0.127	0.257	0.391	0.532	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.819
22	0.127	0.256	0.390	0.532	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.792
23	0.127	0.256	0.390	0.532	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.767
24	0.127	0.256	0.390	0.531	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.745
25	0.127	0.256	0.390	0.531	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.725
26	0.127	0.256	0.390	0.531	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.707
27	0.127	0.256	0.389	0.531	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.690
28	0.127	0.256	0.389	0.530	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.674
29	0.127	0.256	0.389	0.530	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.659
30	0.127	0.256	0.389	0.530	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.646
40	0.126	0.255	0.388	0.529	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.551
60	0.126	0.254	0.387	0.527	0.679	0.848	1.046	1.296	1.671	2.000	2.390	2.660	3.460
120	0.126	0.254	0.386	0.526	0.677	0.845	1.041	1.289	1.658	1.980	2.358	2.617	3.373
∞	0.126	0.253	0.385	0.524	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.291

4.4.4.2. Estudio del test de Student

Sabemos que la media muestral, $\overline{x}$, de una población normal, tiene una distribución $N(\mu, \sigma/\sqrt{n})$, mientras que la variable, $n\, s^2/\sigma^2$, tiene una distribución χ^2 con $n-1$ grados de libertad. De acuerdo con el párrafo anterior, 4.4.4.1., la variable

$$t = \sqrt{n-1}\ \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} : \frac{\sqrt{n}\ s}{\sigma} = \sqrt{n-1}\ \frac{\bar{x}-\mu}{s}$$

tendrá una distribución de Student con n- 1 grados de libertad. Es decir, fijada una probabilidad o coeficiente de confianza, - c%, se verificará

$$P\left(-t_c < \sqrt{n-1}\ \frac{\bar{x}-\mu}{s} \leq t_p\right) = \frac{c}{100} \qquad \text{(I)}$$

o lo que es lo mismo

$$P\left(\bar{x} - t_c \frac{s}{\sqrt{n-1}} \leq \mu < x + t_c \frac{s}{\sqrt{n-1}}\right) = \frac{c}{100}$$

El valor t_c correspondiente nos da, por lo tanto, el intervalo de confianza (figura I-4.33.)

$$\left(\bar{x} - t_c \frac{s}{\sqrt{n-1}}\ ,\ \bar{x} + t_c \frac{s}{\sqrt{n-1}}\right)$$

que nos merece la estimación $\hat{\mu} = \bar{x}$ al nivel c%.

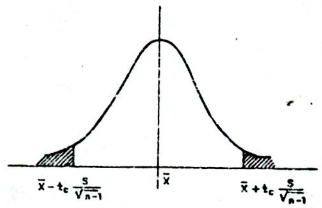


Fig. I-4.33

Pueden considerarse muchos intervalos sin más que escoger dos valores, t_c y t_c', que verifiquen la ecuación (I), pero es preferible siempre escoger un intervalo centrado como se ha hecho.

4.4.5. Indices estadísticos

Los índices estadísticos más utilizados en hidrología - son los siguientes:

Coeficiente de variación

$$C_V = \frac{\sigma_x}{\bar{x}} \qquad \begin{cases} \bar{x} = \dfrac{\Sigma x}{n} \\[2ex] \sigma_x = \dfrac{\Sigma (x-\bar{x})^2}{n-1} \end{cases}$$

Este índice nos da una idea de la irregularidad a efectos comparativos, siendo mayor la irregularidad de la serie --- cuanto mayor es C_V.

Coeficiente de asimetría

$$C_A = \frac{\Sigma (x-\bar{x})^3}{n\ C_V^3}$$

Este coeficiente indica la asimetría de la ley de distribución. Si $C_A > 0$, la distribución se desplaza hacia valores máximos y si $C_A < 0$, la distribución se desplaza hacia valores mínimos.

Error normal

$$e_x = \pm \frac{\sigma_x}{\sqrt{n}}$$

Este índice da el límite probable de error posible en el cálculo de valores medios.

Error total

$$E_x = \frac{\sigma_x}{\sqrt{n}} + \frac{N - n}{N}\ \sigma_x\ \sqrt{1 - r^2}$$

Este índice da el error total posible de la serie total de datos originales y estimados por correlación.

N = número total de datos
n = número de datos originales
r = coeficiente de correlación

4.5. MODELOS HIDROLOGICOS. PROCESOS Y SISTEMAS

Los modelos hidrológicos se consideran como formulaciones matemáticas para simular fenómenos hidrológicos naturales, los cuales son considerados como procesos o sistemas.

Cualquier fenómeno que experimenta continuos cambios, particularmente con respecto al tiempo, puede ser llamado un "proceso". Como prácticamente todos los fenómenos hidrológicos cambian con el tiempo, son "procesos" hidrológicos. Si la oportunidad de ocurrencia de las variables relacionadas en tal proceso se ignora y el modelo se utiliza para seguir cierta ley definitiva, el proceso y su modelo son "determinísticos". Por otra parte, si la ocurrencia de las variables se considera y se introduce el concepto de probabilidad en la formulación del modelo, el proceso y su modelo son "estocásticos o probabilísticos".

Para el proceso probabilístico independiente del tiempo, se ignora la secuencia de ocurrencia de las variaciones relacionadas en el proceso y se admite que el cambio de su ocurrencia sigue una distribución de probabilidad definitiva, en la cual las variables están consideradas como aleatorias puras. Para el proceso estocástico dependiente del tiempo, la secuencia de ocurrencia de las variaciones es observada y las variables pueden ser, aún, aleatorias puras o impuras, y la distribución de probabilidades de las variables puede o no variar con el tiempo. Si en la aleatoria pura, las componentes de las series temporales son independientes entre ellas, constituyen una serie aleatoria. En la aleatoria impura, las componentes -

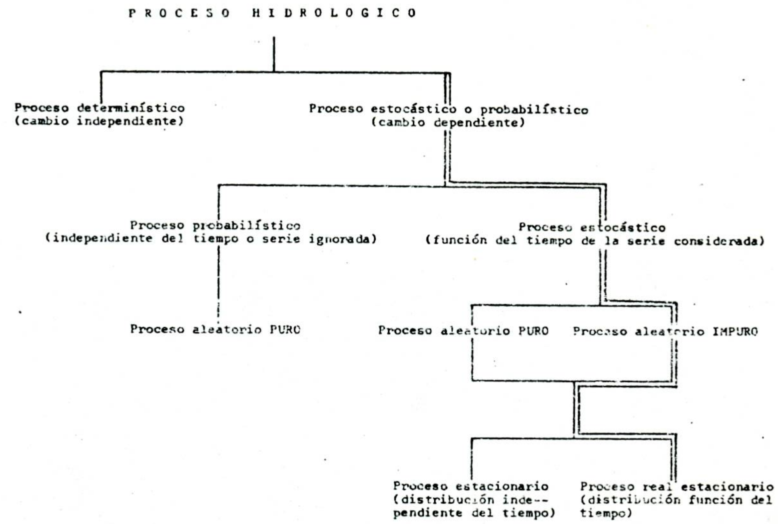
de las series temporales son dependientes entre ellas y están formadas por un componente determinístico y un componente aleatorio puro, constituyendo una serie no aleatoria.

En realidad, todo proceso hidrológico es más o menos estocástico. Ellos toman forma determinística o probabilística solamente para simplificar sus análisis. Hablando matemáticamente, un proceso estocástico es de la familia de las variables aleatorias, x (t), en la cual x es función del tiempo (u otros parámetros) y la variación de x_t se produce a lo largo del tiempo, dentro de un intervalo de tiempo, T.

Cuantitativamente, el proceso estocástico, que puede ser discreto o contínuo, puede ser mostrado continuamente o a intervalos discretos o uniformes de t = 1, 2 ... , y los valores de las muestras forman una serie x_y, x_z, ... , empezando en cierto tiempo y extendiéndose hasta un período de tiempo T. Esta secuencia de valores de la muestra es conocida como una serie temporal, que puede ser discreta o continua (un hidrograma es una serie temporal continua).

La variable aleatoria, x (t), tiene una cierta distribución de probabilidades. Si esta distribución resulta constante a lo largo del proceso, el proceso y las series temporales son llamados "estacionarios". Otras son "no estacionarios". Por ejemplo, un flujo sin cambios significativos en las características de la cuenca hidrográfica o de las condiciones climáticas, para el período de registros, es considerado como series temporales estacionarias. Esto está afectado por la actividad del hombre en la cuenca hidrográfica o los accidentes naturales o la lenta modificación de las lluvias o de las condiciones de escorrentía superficial. En el proceso no estacionario, la complicación matemática es excesiva, por lo que los procesos hidrológicos son tratados como estacionarios.

Para mayor claridad, la clasificación de los procesos hidrológicos se muestra en el esquema de la página siguiente. Se nota que los procesos hidrológicos son procesos que siguen el curso de las líneas gruesas; aunque el proceso siga por las líneas finas, son solamente aproximaciones que pueden ser admitidas con la finalidad de simplificar el análisis.



Un sistema es una agregación o centralización de unidades bajo una forma regular de interacción o independencia. El sistema es llamado DINAMICO si el proceso toma lugar en él. Si el proceso es considerado probabilístico o estocástico, el sistema se dice "estocástico". Consecuentemente, es un sistema determinístico. Posteriormente, el sistema es llamado secuencial si éste consiste en entradas y salidas o cualquier trabajo del flúido (materia, energía o información) conocido como etapa de transición del sistema.