Asignatura: Modelización en Geografía Física

Creative Commons

Reconocimiento – No comercial – Compartir Igual España

CONDICIONES

Usted es libre de

Copiar, distribuir y comunicar públicamente esta obra
hacer obras derivadas

Bajo las condiciones siguientes:

Reconocimiento. Debe reconocer los créditos de la obra de la manera especificada por el autor o el licenciador (pero no de una manera que sugiera que tiene su apoyo o apoyan el uso que hace de su obra).

No comercial. No puede utilizar esta obra para fines comerciales.

Compartir bajo la misma licencia. Si altera o transforma esta obra, o genera una obra derivada, sólo puede distribuir la obra generada bajo una licencia idéntica a ésta.

- Al reutilizar o distribuir la obra, tiene que dejar bien claro los términos de licencia de esta obra.
- Alguna de estas condiciones puede no aplicarse si se obtiene el permiso del titular de los derechos de autor
- Nada en esta licencia menoscaba o restringe los derechos morales del autor.

Índice

1 Sistemas, Modelos y Medio Ambiente **9**

1.1 Concepto de sistema . . . 9
1.2 Naturaleza de la experimentación científica . . . 9
1.3 Situación en las Ciencias Ambientales . . . 10
1.4 La modelización como solución . . . 11
1.5 Tipos de modelos . . . 14
1.6 Tipos de modelos matemáticos . . . 16
1.7 Desarrollo tecnológico y modelización . . . 20

2 Introducción a R **23**

2.1 Sobre R . . . 23
2.2 Primeros pasos con R . . . 24
2.2.1 Ayuda en R . . . 26
2.2.2 Expresiones en R . . . 26
2.2.3 Variables . . . 28
2.2.4 Vectores . . . 29
2.2.5 Funciones . . . 31
2.2.6 Matrices y data.frames . . . 32
2.2.7 Funciones genéricas . . . 33
2.2.8 Funciones de usuario . . . 34
2.3 Importación y exportación de datos . . . 34
2.4 Gráficos con R . . . 35
2.4.1 Modificación del aspecto de puntos y lineas . . . 39
2.4.2 Guardar los gráficos . . . 40

3 Estadística descriptiva. Análisis de muestras **43**

3.1 Introducción . . . 43

3.2 Estadística univariante . . . 44

3.2.1 Gráficos . . . 44

3.2.2 Estadísticos . . . 44

3.2.3 Con R . . . 50

3.2.4 Ejercicios . . . 50

3.3 Estadística bivariante . . . 51

3.3.1 Gráficos bivariantes . . . 51

3.3.2 Correlación paramétrica . . . 51

3.3.3 Correlación no paramétrica . . . 52

3.3.4 Con R . . . 53

3.3.5 Ejercicios . . . 54

4 Estadística inferencial **57**

4.1 Probabilidad . . . 59

4.2 Inferencia a partir de los estadísticos descriptivos . . . 62

4.2.1 Con R . . . 62

4.2.2 Error estándar de la media . . . 66

4.2.3 El error estándar de la desviación típica . . . 67

4.2.4 El error estándar del coeficiente de sesgo . . . 67

4.2.5 Errores estándar de un conjunto de proporciones . . . 69

4.2.6 Aplicación de los errores típicos . . . 69

4.3 Inferencia acerca de la correlación . . . 70

4.4 Contraste de hipótesis . . . 71

4.4.1 Contraste de hipótesis respecto a la media . . . 72

4.4.2 Contraste de hipótesis respecto a l amedia con datos pareados . . . 73

4.4.3 Con R . . . 73

4.5 Recapitulación . . . 75

5 Inferencia acerca de la distribución **77**

5.1 Modelos de distribución para variables discretas 78

5.2 Modelos de distribución para variables continuas 83

5.3 Con R 99

5.4 ¿Cómo estimar los parámetros de una función de distribución? 100

5.5 ¿Cual es la distribución que mejor representa mis datos? 102

5.5.1 El gráfico Q-Q 103

5.5.2 Test χ^2 103

5.5.3 El test de Kolmogorov-Smirnov 106

6 Relación entre variables **107**

6.1 Análisis de regresión 107

6.1.1 Resultados de un análisis de regresión 109

6.1.2 Con R 110

6.2 Regresión y correlación múltiple 112

6.2.1 Con R 113

6.3 Análisis de varianza 113

6.3.1 Con R 116

6.3.2 Ejemplos 117

7 Programación **119**

7.1 Introducción a la computación 119

7.2 Lenguajes de programación 120

7.3 Fases en el desarrollo de un programa 121

7.4 Elementos de programación 121

7.4.1 Variables y operadores 121

7.4.2 Entrada y salida 122

7.4.3 Estructuras de control: Bucles 122

7.4.4 Estructuras de contro: Toma de decisiones 124

7.4.5 Funciones definidas por el usuario 125

7.5 Estructura de un programa 126

8 Modelos empíricos **129**

8.1 Fases en la construcción de un modelo empírico 130

8.1.1 Identificación 130

8.1.2 Calibración 130

8.1.3 Validación y Verificación 130

8.1.4 Análisis de sensibilidad 131

8.2 El modelo de erosión de Thornes(1990) 132

8.3 Cuestiones 135

8.4 Modelos para generar series temporales de variables 135

8.4.1 Series anuales de variables climáticas 135

8.4.2 Series mensuales de variables climáticas 136

8.4.3 Simulando un episodio de precipitación 138

9 Modelos conceptuales **141**

9.1 Método racional en hidrología 141

9.1.1 Agregado y estático 142

9.1.2 Semidistribuido y estático 142

9.1.3 Semidistribuido y dinámico 143

9.1.4 El código 144

9.1.5 Cuestiones 145

10 Modelos de balance de materia y energía **147**

10.1 Un modelo de infiltración basado en la ecuación de Green-Ampt 147

10.1.1 El código 149

10.1.2 Ejercicios 150

10.2 Un modelo de balance hídrico 150

10.2.1 El código 150

10.2.2 Ejercicios 151

10.3 El mundo de las margaritas 152

10.3.1 El código 152

10.3.2 Cuestiones 153

11 Modelos de base física **157**

11.1 Modelos hidrológicos distribuidos de tipo físico: Un modelo de onda cinemática 158

11.1.1 El código 160

11.1.2 Ejercicios 161

Tema 1

Sistemas, Modelos y Medio Ambiente

1.1 Concepto de sistema

Uno de los conceptos más ampliamente utilizados en la investigación científica es el de **sistema**. La definición más habitual de **sistema** es la debida a Chorley y Kennedy (1971) que definieron sistema como *un conjunto estructurado de componentes y variables que muestran relaciones entre ellos y operan en conjunto como un todo complejo de acuerdo con unas pautas observadas*.

Un sistema se percibe como algo que posee una entidad que lo distingue de su entorno, aunque mantiene una interacción con él. Esta identidad permanece a lo largo del tiempo y bajo entornos cambiantes.

En Ciencias de la Tierra y Ambientales se trabaja con diversos conceptos derivados de este como son **ecosistema**, **geosistema**, **sistema fluvial**, etc.

1.2 Naturaleza de la experimentación científica

Tradicionalmente se ha considerado que la **investigación científica** se desarrolla a través de la acumulación de observaciones del comportamiento de los sistemas estudiados en circunstancias naturales o manipuladas a través de un experimento. Estas observaciones permiten generar y contrastar hipótesis acerca de la estructura y función del sistema objeto de estudio para incrementar los conocimientos acerca del mismo.

Un **experimento** puede definirse como la obtención de una serie de variables de uno o varios individuos, previamente seleccionados de una población, con el objeto de comprobar una hipótesis o desarrollar una teoría. Ello implica un control absoluto de todas las variables y factores vinculadas con el experimento.

En sentido estricto sólo puede hablarse de experimentos en aquellas ciencias como la física o bioquímica, en las que resulta fácil aislar los elementos que se quieren controlar. Los experimentos en física se dividen en una serie de pasos:

1. Observación de un efecto

2. Formulación de hipótesis acerca del efecto observado
3. Medición de las variables dependientes e independientes
4. Modificación controlada de los factores independientes para producir el efecto deseado
5. Replicación de (3) para permitir la falsación [1] de (2)
6. Derivación de leyes que expliquen la relación entre las variables
7. Establecimiento del rango de aplicabilidad de estas leyes

Por ejemplo, un bioquímico puede observar que al incluir sacarosa en la alimentación de un cultivo bacteriano este crece más rápido (1). Esta observación se convierte en la hipótesis de que la sacarosa es un buen alimento para los cultivos bacterianos (2), debiéndose controlar la variable dependiente (crecimiento) e independiente (dosis de sacarosa) así como otros factores que pudieran modificar la relación como temperatura, humedad, iluminación, etc. (3 y 4). La replicación del experimento permitirá determinar si se cumple la hipótesis en todas las condiciones (5) y pasa a convertirse en ley (6) o si existen determinadas condiciones ambientales (valores de los factores) para los cuales no se cumple (7).

1.3 Situación en las Ciencias Ambientales

En el conjunto de las ciencias de la Tierra y medioambientales, incluyendo la Geografía Física y todas sus ramas, la observación de efectos y el establecimiento de hipótesis resulta más difícil debido a todo un conjunto de factores:

- **Complejidad** del fenómeno estudiado. Los procesos que actúan sobre el territorio se caracterizan por su carácter tridimensional, su dependencia del tiempo y complejidad. Esta complejidad incluye comportamientos no lineales, componentes estocásticos, bucles de realimentación a diferentes escalas espaciales y temporales haciendo muy complejo, o incluso imposible, expresar los procesos mediante un conjunto de ecuaciones matemáticas. Las causas de esta complejidad son variadas:
  - Las **relaciones no lineales** implican que pequeñas causas puedan tener como consecuencia grandes efectos.
  - **Discontinuidad y bimodalidad**, existencia de diversos estados de equilibrio.
  - **Histeresis**, los procesos no son exactamente reversibles.
  - **Divergencia**, existencia de varios efectos posibles para una misma causa.
  - Imbricación de **múltiples causas y efectos**.

[1] Falsación en el sentido dado por K.Popper al término, es decir la posibilidad de demostrar que la hipótesis es falsa. Según este autor, toda teoría científica debe incluir esta posibilidad para poder considerarse como tal

  - El flujo de materia o energía no se traslada de un componente a otro del sistema sino que puede hacerlo de uno a varios o viceversa.
  - Las relaciones de **retroalimentación** (los efectos son causas de sus causas) son importantes y hacen que un sistema sea más complejo y su comportamiento difícil de predecir. Puede existir retroalimentación positiva (**homeorhesis**) que una vez iniciado el cambio tienden a mantenerlo o incrementarlo, o negativa (**homeostasis**) que tienden a compensar el cambio eliminándolo.

- **Imposibilidad de control**. En otras ciencias (física, química, etc.) es posible mantener los sistemas estudiados en condiciones controladas de laboratorio, en las ciencias ambientales este enfoque resulta imposible. Cualquier intento de llevar una porción del sistema al laboratorio implica una mutilación del mismo y la modificación total de las condiciones de contorno. La opción intermedia de los campos y parcelas experimentales por un lado suponen también un cierto aislamiento de la porción estudiada respecto al resto del sistema y por otro no se consigue el grado de control que se tiene en un laboratorio.
- Muy relacionado con el anterior problema está la **irrepetibilidad de las mediciones**, por ejemplo no podemos volver a medir la lluvia caída un día concreto. Por otro lado los procedimientos de medición suelen suponer la alteración del objeto de estudio durante la medición y tras los experimentos.
- Problemas de **escala** tanto espacial como temporal. Por una parte el problema es que en muchos casos el investigador sólo tenga acceso a una pequeña parte del continuo espacio-temporal en el que ocurren los procesos a estudiar, y que la visión que se puede tener de estos cambia con la escala. Por otro lado un mismo sistema puede comportarse de formas diferentes a diferentes escalas.
- Los sistemas ambientales son **multicomponente**, requieren comprender diversos aspectos geológicos, climáticos, hidrológicos, ecológicos, etc. y su imbricación. Por tanto es imposible que un sólo especialista pueda abordar el estudio de estos sistemas que deben ser necesariamente objeto de una investigación **multidisciplinar**.

Estos problemas, si bien son de difícil solución, no son tan graves como los que aparecen en ciencias sociales en las que el comportamiento de uno solo de los individuos estudiados es posiblemente más complejo que cualquiera de los sistema estudiados en ciencias ambientales.

La investigación en ciencias ambientales tiene por otra parte un proposito fundamentalmente práctico dentro del cual es fundamental hacer prospectivas de futuro (planes de conservación), entender los impactos de acontecimientos que aún no han sucedido (evaluación de riesgos naturales), o evaluar los impactos que puede tener la activiad humana (análisis de impacto ambiental). Para abordar el estudio de todo este tipo de problemas en unos sistemas que se caracterizan por el grado de complejidad antes mencionado, es imprescindible ser capaces de hacer una simplificación racional de los mismos.

1.4 La modelización como solución

Debido a la dificultad de llevar a cabo experimentos auténticos que cumplan con los criterios antes mencionados y que respondan a las necesidades prácticas de la investigación sobre sistemas ambientales se ha propuesto una

amplia gama de modos de trabajo que relajan las estrictas condiciones que debe cumplir un experimento.

Una de estas lineas es el estudio de los sistemas ambientales mediante modelos. Un **modelo** es una *representación simplificada de una realidad compleja de forma que resulte adecuada para los propósitos de la modelización*. Esta simplificación se basa en una serie de asunciones acerca de como funciona un sistema que no son totalmente válidas pero permiten representar el sistema de forma más sencilla. Puesto que la validez del modelo depende de la validez de las asunciones, es importante que estas sean entendidas y establecidas de forma explícita. La necesidad de decidir que aspectos de la realidad deben incluirse en el modelo hace que este refleje no sólo la realidad del sistema sino también la percepción que el investigador tenga del mismo o lo que es lo mismo las teorías que considera más adecuadas para representarlo. De este modo, un modelo se convierte de forma secundaria en un vehículo de transmisión de conocimiento.

En esa versión simplificada y fácil de manejar de la realidad que constituye un modelo, podremos estudiar como diversas políticas, acontecimientos imprevistos o nuevas actividades afectan a los elementos del sistema representados, bajo la hipótesis de que los resultados serán similares en la realidad. Se experimenta por tanto con el modelo no con el sistema de manera que se solventan muchos de los problemas prácticos mencionados más arriba.

Por ejemplo, un modelo a escala de un cauce (figura 1.1) consiste en una maqueta que representa dicho cauce en tamaño reducido, utilizando materiales que también aparecen a escala (los grandes bloques pueden simularse con cantos, los cantos con arena gruesa, la arena con limo, etc.[2]). Si dejamos entrar al modelo diferentes caudales, podremos observar las consecuencias (areas inundadas, erosión, etc.) en todo el modelo y asumir que el sistema fluvial real se comportará igual.

Para lograr todos estos objetivos, la construcción de un modelo debe venir precedida de:

1. Una **definición clara de los objetivos del modelo**; entre los objetivos habituales en modelización destacan:

   - **Simulación**:
     - Explorar situaciones (escenarios) en las que no contamos con datos empíricos
     - Interpolar entre medidas para estimar el valor de la variable
     - Estimar el valor de una variable a partir de otras
   - **Predicción**: Extrapolar más allá del espacio-tiempo donde tenemos medidas
   - **Incremento del conocimiento**:
     - Acerca del funcionamiento de los sistemas:
     - Acerca de cuales son los parámetros, variables, relaciones, procesos, estructuras y escalas importantes
   - **Apoyo a la investigación científica**:
     - facilitando la integración multidisciplinar

[2] En todo caso hay que tener en cuenta que el gran problema de estos modelos fluviales son las grandes diferencias entre arenas, limos y arcillas en cuanto a su comportamiento físico-químico

Figura 1.1: Modelo a escala de un cauce

  - proporcionando un *laboratorio virtual*
  - facilitando la comunicación de la investigación y sus resultados
- **Gestión diária**, mediante la simulación por adelantado de los caudales previstos en función de la lluvia. Los resultados deben obtenerse en un tiempo razonablemente breve y a bajo coste, aunque el modelo no sea el mejor posible.

2. Una **identificación adecuada de los elementos y procesos involucrados** en el sistema a modelizar y relevantes para alcanzar los objetivos planteados;

3. **Determinación de la escala espacial y temporal más adecuadas**. Cuando se trabaja en modelización el concepto de escala incluye tanto la extensión del marco espacio-temporal del modelo como la resolución del mismo, es decir los intervalos espaciales y temporales en que se va a subdividir ese marco. La elección del marco espacial implica establecer una frontera entre el sistema a estudiar y su entorno que no no será objeto de estudio;

4. Pero puesto que ningún sistema es cerrado, en necesario determinar que **flujos de materia, energía o información** que se producen entre el sistema y su entorno y son relevantes para los objetivos del modelo.

Por otro lado, en la construcción de un modelo hay que tener siempre presente el **principio de parsimonia**, aunque el añadir nuevos parámetros y variables pueda hacer al modelo teóricamente más potente, esta adición

esta siempre sujeta a una *ley de rendimientos decrecientes* en cuanto que la mejoras en el modelos serán cada vez más pequeñas cuanto más elementos introduzcamos. Además un modelo con un gran número de parámetros resulta más difícil de entender, calibrar o validar.

La modelización se ha convertido en un procedimiento de trabajo frecuente en aquellas situaciones en las que:

- Ensayar sobre sistemas reales puede resultar muy costoso o puede llevar a la destrucción de los mismos
- Puede ser interesante alterar las escalas de tiempo para estudiar procesos que en la naturaleza ocurren a largo plazo mediante su aceleración.

La modelización resulta por otro lado una herramienta de gran importancia para el desarrollo de la investigación científica. Un modelo matemático del funcionamiento de un sistema ambiental, integra diversas ecuaciones que modelizan a su vez aspectos concretos del mismo, cada una de estas ecuaciones representa por su parte una teoría, una hipótesis acerca del funcionamiento de dicho aspecto. Los modelos se convierten de este modo en *juegos de construcción* en los que integrar diferentes teorías científicas, estudiar su correspondencia con la realidad, así como sus interacciones.

Sirven por otro lado para representar, transmitir y ayudar a comprender teorías científicas. Aunque el lenguaje matemático y computacional sea complejo y rígido permite gran flexibilidad para representar una teoría de forma objetiva con lo que se facilita su comprensión a los demás. Esta facilidad para la transmisión de ideas y conocimientos favorece además el trabajo interdisciplinar. Cada miembro de un equipo multidisciplinar puede expresar el conocimiento que desde su campo se tiene sobre el sistema objeto de estudio como un componente de un modelo y entender fácilmente el conocimiento aportado por otros especialistas sobre aquellos aspectos con los que su componente debe interaccionar.

Un buen modelo, suficientemente validado, puede servir finalmente como herramienta para la gestión diaria de los sistemas ambientales. Para ello se le debe añadir una interfaz adecuada para que usuarios no expertos puedan predecir las consecuencias de acciones concretas o cuales son las actuaciones más adecuadas para conseguir determinado objetivo.

1.5 Tipos de modelos

- **Conceptual:** responde a una descripción del sistema y su funcionamiento utilizando el lenguaje humano. Suele ser la fase previa al desarrollo de cualquier modelo. Implica seleccionar que parámetros[3] y variables son los más relevantes, así como las relaciones de causalidad positiva o negtiva (a mayor precipitación mayor escorrentía, a mayor albedo menor radiación neta) o control que se establecen entre ellos.
- **Icónico:** se basa en la representación de los componentes dle sistema mediante símbolos, en las figuras 10.1 y 10.3 aparecen dos ejemplos. Los modelos icónicos compelmentan a los conceptuales.Se han desarrollado diversos conjuntos de símbolos para crear modelos icónicos, uno de los más utilizados son los **diagramas de Forrester** (figura 1.4 que distinguen entre:

[3]Los parámetros son magnitudes que se asumen constantes a los efectos del modelo

  - **Componentes** o almacenamientos de materia o energía
  - **Flujos** de materia o energía
  - **Tasas** que controlan la actividad de los flujos
  - **Fuentes** que son componentes externos al sistema desde los que entran los flujos al mismo
  - **Sumideros** que son componentes externos al sistema desde los que salen los flujos al mismo

- **Físico:** basado en prototipos construidos para estudiar el sistema. Existen dos tipos:
  - **Analógico**: Se basa en la analogía existente entre dos sistemas físicos diferentes uno, el que nos interesa estudiar, y otro mucho más sencillo de estudiar que actúa como modelo. Por ejemplo se han utilizado sistemas electrónicos como analogías del comportamiento de los acuíferos en los que la transmisividad del acuífero se simulaba en el modelo añadiendo resistencias. También se han utilizado modelos analógicos para estudiar los flujos de agua en el interior de una planta.
  - **a escala**: Se trata de modelos reducidos del sistema que se estudia. Se han utilizado mucho en hidrología, plantean la dificultad del cambio de comportamiento de los materiales con el cambio de escala. Por ejemplo las arena podría sustituirse por arcilla pero esta presenta comportamientos cohesivos que no tiene aquella.

- **Matemático:** Son los más utilizados actualmente y se basan en la representación del estado de los componentes de un sistema y los flujos entre ellos mediante un conjunto de ecuaciones matemáticas. Pueden ir desde un conjunto de ecuaciones simples a programas complejos que incluyen una gran cantidad de ecuaciones y reglas y que, por tanto, requieren un ordenador para su resolución. La clasificación de los modelos matemáticos resulta bastante compleja ya que hay que tener en cuenta diversas consideraciones.

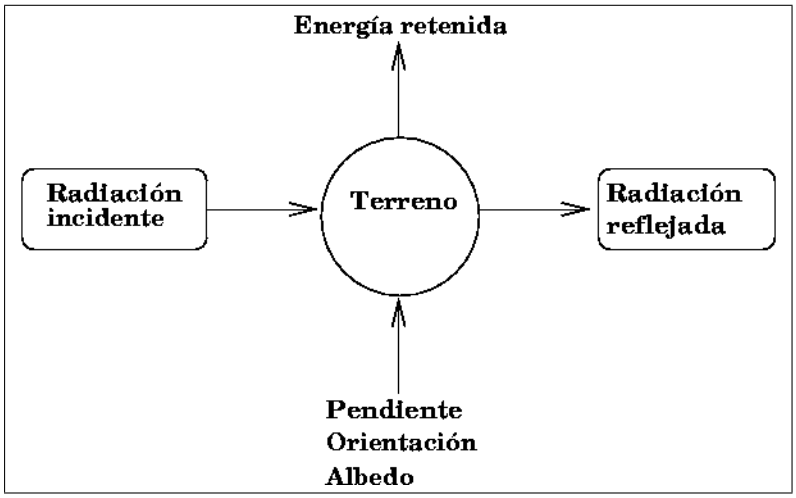


Figura 1.2: Modelos de radiación

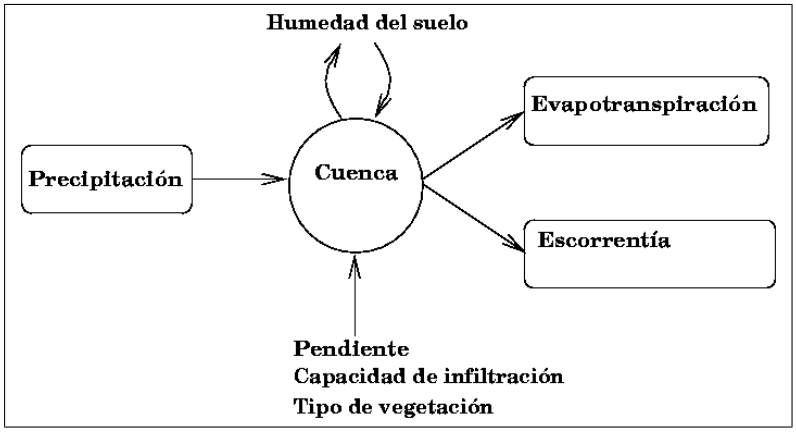


Figura 1.3: Modelos de procesos en una cuenca

1.6 Tipos de modelos matemáticos

Una clasificación de los modelos matemáticos podría basarse en una serie de características dicotómicas:

- **Empíricos o basados en principios físicos**

  El carácter empírico o físico constituye la característica fundamental de un modelo. un **modelo empírico** se basa en relaciones estadísticamente significativas entre variables. Las ecuaciones que describen un modelo estadístico no son por tanto físicamente o dimensionalmente consistentes ni universales, ya que en rigor sólo son válidas para el contexto espacio-temporal en el que se obtuvieron y calibraron. Se caracterizan por un alto poder predictivo pero una escasa capacidad explicativa, es decir reproducen el funcionamiento del sistema razonablemente bién pero no permiten saber por que el sistema funciona así. Los modelos estadísticos se conocen también como **modelos de caja negra** ya que no permiten descubrir el funcionamiento interno del sistema.

  Un **modelo físico** se basa en las leyes físicas que rigen los procesos, se denominan, por contraposición, **modelos de caja blanca**. Se trata de modelos en los que las transferencias de materia y energía entre sus componentes se rigen mediante ecuaciones físicas y que además cumplen las leyes de conservación de la materia y la energía, tanto para el conjunto del modelo como para cada uno de los submodelos.

  Una posibilidad intemedia son los **modelos de caja gris** o **conceptuales**. Se trata de modelos en los que el sistema se descompone en una serie de componentes que se resuelven como modelos empíricos pero cuya integración se basa en principios físicos o al menos en cierto conocimiento a priori de como funciona el sistema. Dentro de los modelos conceptuales pueden incluirse los **modelos de balance** en los que se estudian los flujps de materia y energía a través de una serie de componentes respetando los principios de conservación de materia o energía cuyas sumas totales no deben variar.

  Para poder construir un modelo físico es necesario un alto conocimiento acerca de como funciona el sistema a modelizar. El modelo resultante permite transformar unas variables de entrada en variables de

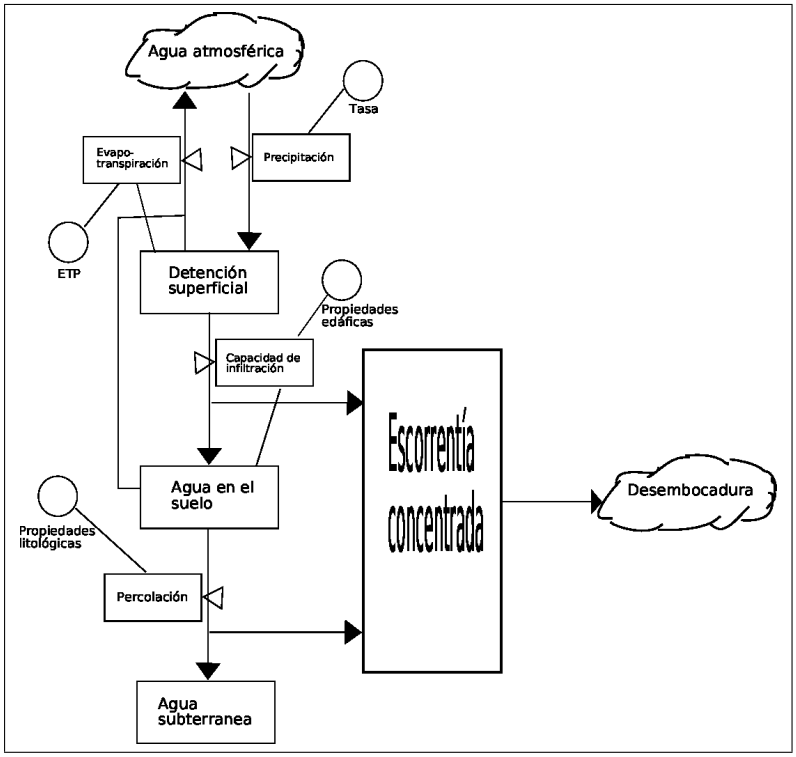


Figura 1.4: Diagrama de Forrester de un modelo de procesos en una cuenca

salida. En el caso de los modelos empíricos la falta de conocimiento acerca del sistema se compensa con datos de calidad y en cantidad suficiente de las variables de entrada y de salida. A partir de estos datos podemos construir un modelo empírico que nos permita, posteriormente, obtener las variables de salida a partir de un nuevo conjunto de valores para las variables de entrada.

- **Estocásticos o deterministas**

  Los primeros incluyen generadores de procesos aleatorios dentro del modelo que modifican ligeramente algunas de las variables. De esta manera, para un mismo conjunto de datos de entrada, las salidas no serían siempres las mismas. La distinción ente modelos deterministas o estocásticos se confunde a veces con la anterior, relacionando equivocadamente modelos estocásticos con empíricos y deterministas con físicos. En realidad un modelo determinista es aquel en el que dado un conjunto de de parámetros y variables de entrada va a producir siempre el mismo conjunto de variables de salida. En el caso de un modelo estocástico los valores de las variables de salida van a a variar de unas ejecuciones del modelo a otras ya que se deja intervenir al azar.

  Las razones por las que puede ser interesante introducir cierto grado de aleatoriedad en el comportamiento de un modelo son diversas:

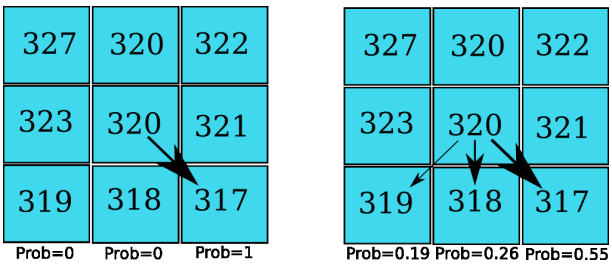


Figura 1.5: Distribución de flujo entre celdillas de un MDE de forma determinista y estocástica

- Existencia de procesos realmente estocásticos dentro del sistema
- Existencia de procesos aparentemente estocásticos debido a nuestra falta de conocimiento
- Errores aleatorios en la medición de las variables con lo cual sabemos que los valores reales pueden oscilar más o menos entorno al valor medido
- Existencia de procesos muy complejos que es preferible modelizar como estocásticos

La introducción de un componente estocástico en un modelo puede, por otro lado, conseguirse mediante:

- Utilización de generadores aleatorios de series de las variables de entrada partiendo de un modelo adecuado de distribución de probabilidad (ver sección 5)
- Utilización de generadores aleatorios para dar valores a los parámetros del modelo y a su distribución espacial, de esta manera se evita el problema que supone utilizar parámetros estimados con cierto grado de incertidumbre. Por ejemplo al incluir la capacidad de infiltración del suelo en un modelo hidrológico podemos utilizar siempre los mismos valores o permitir que varie al azar de unas ejecuciones a otras.
- Cuando en un modelo las salidas de un componente pueden dirigirse a varios componentes distintos y no es fácil determinar a cual o en que cantidad, pueden determinarse las cantidades al azar. Por ejemplo en un modelo hidrológico basado en la rasterización de una cuenca, la transferencia de agua de una celdilla a las celdillas aguas abajo puede hacerse de modo determinista (siempre igual) o aleatorio (puede variar), ver la figura 1.5.

La introducción del componente estocástico permite además comprobar como se comportaría el modelo para diferentes conjuntos de parámetros o valores de las variables de entrada. De este modo en lugar de obtener un único resultado a partir de un conjunto fijo de parámetros y variables de entrada, obtendremos un conjunto de resultados a partir de varios conjuntos verosímiles de parámetros y variables. De este modo no tenemos por que conformarnos con un valor esperable sino que tendremos un rango de variación dentro del cual estarán los resultados esperables. Por ejemplo en el caso de estimación del riesgo de inundación resulta muy difícil determinar cual será la altura máxima de la lámina de agua. Los modelos estocásticos permitirían obtener una distribución de probabilidades de altura de agua, lo que sería más interesante de cara a la planificación del territorio.

- **Agregados o distribuidos**

  En el caso de los **modelos agregados**, toda el área de estudio se considera de forma conjunta, por ejemplo una cuenca hidrográfica. Se tiene un único valor para todos los parámetros del modelo. El modelo predice unas salidas para las entradas aportadas sin informar de lo que ocurre dentro del sistema. Un verdadero modelo distribuido no tienen en cuenta sólo las variación espacial de los procesos (como si se tratara de muchos modelos agregados yuxtapuestos) sino también las transferencias de materia y energía en el espacio.

  En un **modelo distribuido**, tendremos el área de estudio dividida en porciones cada una de ellas con su propio conjunto de parámetros y sus propias variables de estado. Cada porción recibe un flujo de materia y energía de algunas de sus vecinas que a su vez reemite a otras.

  Una tercera posibilidad son los **modelos semidistribuidos** que se construyen a partir de la yuxtaposición de diversos modelos agregados, por ejemplo diversas subcuencas de una cuenca hidrográfica. Otra posibilidad a menudo explorada en hidrología es dividir el área de trabajo en *Unidades de Respuesta Hidrológica*. Se trata de segmentos de ladera homogeneos en cuanto a su pendiente, orientacion, litología y uso a los que se asume una respuesta hidrológica única. En un modelo semidistribuido las diferentes unidades generan sus propias salidas de forma agregada pero aparecen entradas y salidas de unas a otras.

  La incorporación de la componente espacial en los modelos resulta bastante compleja. Si se opta por un modelo distribuido es necesario establecer un **modelo de datos espaciales** que permita asignar valores de los parámetros y las variables de estado a los diferenes puntos del área de estudio. Puede tratarse de **distribuciones de puntos**, de mallas **raster** o de redes irregulares de triángulos (**TIN**). Si se trabaja con modelos agregados o semidistribuidos hay que codificar, además, los límites de las diferentes unidades. Todos estos proceso son más complejos de lo que pudiera parecer a primera vista y la manera más eficiente de hacerlo es mediante un **Sistema de Información Geográfica**.

- **Estáticos o dinámicos**

  Se refiere a la forma en que se trata el tiempo. Los modelos estáticos dan un resultado agregado para todo el período de tiempo considerado este puede ser por ejemplo un caudal medio o un caudal punta. Los modelos dinámicos devuelven las series temporales de las variables consideradas a lo largo del período de estudio. Por ejemplo podemos considerar un modelo estático de cuenca en el que el caudal medio ($\bar{Q}$) depende de la precipitación total (P) y de los parámetros de la cuenca (θ):

  $$\bar{Q} = f(P, \theta) \tag{1.1}$$

  o un modelo dinámico en el que el caudal en cada intervalo de tiempo considerado (Q_{t+1}) depende de la precipitación no sólo en dicho intervalo de tiempo sino tambien en los intervalos anteriores y de los parámetros de la cuenca:

  $$Q_t = f(P_t, P_{t-1}, P_{t-2}, ..., \theta) \tag{1.2}$$

Recuerda que los parámetros se distinguen de las variables en que aquellos son invariantes a la escala espacio-temporal del modelo. Las **variables de entrada y salida** representan flujos de materia y enegía desde y hacia

el interior del sistema (precipitación y caudal por ejemplo). Las **variables de estado** representan cambios en la cantidad de materia y energía disponible (humedad del suelo). La distinción entre variables y parámetros depende de la escala, espacial y temporal, del modelo.

Otras variables de importancia en un modelo dinámico son las **condiciones iniciales** que representan el valor de las variables de estado al comienzo de la simulación y las **condiciones de contorno** que son los valores de variables relevantes externas al sistema pero que influyen sobre este (entre ellas las variables de entrada).

En definitiva, un sistema natural recibe entradas de materia y energía de su entorno que devuelve a dicho entorno con ciertas modificaciones. Entre estas modificaciones cabe destacar:

- Desplazamiento en el espacio
- Modulación en el tiempo de los flujos

1.7 Desarrollo tecnológico y modelización

Hasta los años 50 del siglo pasado, la falta de capacidad de cálculo imposibilitaba la aplicación de modelos, lo que no impidió el desarrollo de los mismos. De hecho gran parte de las leyes físicas que rigen los procesos ambientales se desarrollaron en los siglos XVIII y XIX. Debido a la dificultad en los calculos se tabulaban los resultados de las ecuaciones más complejas, pero estas no se podían interrelacionar.

Durante la segunda mitad del siglo XX, el formidable desarrollo tecnológico experimentado ha permitido atacar problemas y planteamientos teóricos que antes resultaba imposible.

- Enormes incrementos en la capacidad de computación junto a una no menos enorme disminución de sus costes;
- Métodos de medición en continuo, *data loggers*;
- Desarrollo de métodos de muestreo;
- Desarrollo de técnicas basadas en indicadores (en modelos en definitiva) para medir variables ambientales;
- Desarrollo de técnicas computacionales de análisis de datos.

El resultado de estos avances ha sido por un lado la mejor calidad de los estudios llevados a cabo y en segundo lugar la necesidad de una mayor cualificación técnica para llevarlos a cabo.

Desde el punto de vista de la modelización, estos adelantos han permitido la **simulación por ordenador**, es decir la resolución mediante un ordenador de un conjunto de ecuaciones que corresponden a modelos básicos, integradas formando un modelo matemático y reorganizadas de forma algorítmica mediante un programa informático.

En 1982 J.N.F. Jeffers escribía *La modelización es un modo de resolución de los problemas científicos que ha crecido rápidamente en los últimos años (...) estimulado por las grandes posibilidades que dan los ordenadores (...) más baratos, más pequeños y cada vez más potentes*. En 2008 se ha llegado al punto en que cualquier ordenador doméstico es capaz de ejecutar modelos relativamente sofisticados varias veces, de manera que la modelización ha dejado de ser un *oráculo* costoso al que acudir en busca de "la respuesta" sino que nos permite consultar diversas respuestas, bien modificando el modelo o modificando ligeramente los valores de parámetros y variables para, de este modo, obtener en lugar de un valor un rango de valores más probables.

Los modelos se han convertido así en una herramienta de uso rutinario en apoyo a la labor científica o de gestión. Resultan especialmente útiles en el manejo de sistemas complejos.

Ejercicios

- Escoge un problema ambiental y trata de representarlo mediante modelos conceptuales e icónicos resaltando:
  - Tipo de modelo más adecuado (estático-dinámico, agregado-distribuido, determinista-estocástico, físico-empírico)
  - Variables y parámetros a tener en cuenta
  - Tipo de relaciones entre las variables

Tema 2

Introducción a R

2.1 Sobre R

R es un sistema informático para realizar análisis estadísticos y gráficos que permite el desarrollo de modelos matemáticos. Tiene una naturaleza doble de programa y lenguaje de programación y es considerado como un *dialecto* del lenguaje S, desarrollado en los laboratorios de la compañía AT&T Bell. De este lenguaje se han realizado varias implementaciones, algunas de ellas comerciales, la más conocida es S–plus.

La utilización de R se ha extendido de forma creciente y los investigadores lo han seleccionado por sus ventajas, como puede comprobarse en el creciente número de trabajos publicados en la revista *Journal of Statistical Software*, publicada por la *American Statistical Association*[1]. Entre estas ventajas destaca:

- se trata de un programa de distribución libre
- puede utilizarse con distintos sistemas operativos
- adecuación del lenguaja al manejo y análisis de los datos
- sencillez con la que pueden combinarse los distintos análisis estadísticos
- gráficos de alta calidad: visualización de datos y producción de gráficos[2]
- facilidad para incorporar nuevos procedimientos a los proporcionados por el sistema base
- muy completa documentación.
- la comunidad de R es muy dinámica, con gran crecimiento del número de paquetes, e integrada por cienífícos de gran renombre

[1] http://www.jstatsoft.org

[2] http://addictedtor.free.fr/graphiques/thumbs.php?sort=votes

- hay extensiones específicas a nuevas areas como bioinformática, sistemas de información geográfica, geoestadística y modelos gráficos
- es un lenguaje orientado a objetos
- se parece a Matlab y a Octave, y su sintaxis recuerda a C/C++

Sintetizando, puede describirse a R como un entorno integrado para trabajar con el lenguaje S, que proporciona:

- Un conjunto coherente y extensivo de instrumentos para el análisis y el tratamiento estadístico de datos.
- Un lenguaje para expresar modelos estadísticos y herramientas para manejar modelos lineales y no lineales. Las expresiones componen de operadores, valores numéricos, valores lógicos y funciones.
- Utilidades gráficas para el análisis de datos y la visualización en cualquier estación gráfica o impresora.
- Un eficiente lenguaje de programación orientado a objetos, que crece fácilmente merced a la comunidad de usuarios.

El único inconveniente que se le puede achacar es la dificultad asociada a un lenguaje muy rico, aunque existen varias formas de ayuda acerca del lenguaje, y la necesidad de una cierta experiencia en el uso de sus comandos.

Puede encontrarse abundante información sobre el funcionamiento de R, así como los ejecutables para su instalación, en la página principal del proyecto[3]. Se han publicado numerosos textos de tratamiento y análisis de datos con R[4]. En la página de wikipedia[5] pueden encontrarse más referencias.

La instalación de R es relativamente sencilla y varia de un sistema operativo a otro, adecuandose a los procedimientos habituales en cada uno de ellos.

2.2 Primeros pasos con R

Para arrancar el programa, los usuarios de un sistema winXX utilizarán el menú del sistema, tal como muestra la figura 2.1.

Para usuarios de sistemas Linux bastará con invocarlo desde la línea de comandos, escribiendo:

```
usuario@servidor:directorio$ R
```

El sistema entrará en el programa y ofrecerá información básica sobre él y quedará a la espera de recibir una orden[6]:

[3] http://www.r-project.org
[4] http://www.r-project.org/doc/bib/R-books.html
[5] http://en.wikipedia.org/wiki/R_programming_language
[6] esta presentación puede variar ligeramente de una versión a otra del programa.

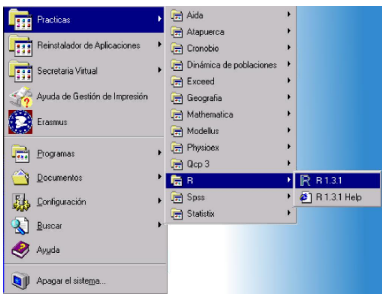


Figura 2.1: R en el menú de inicio de windows

```
R : Copyright 2005, The R Foundation for Statistical Computing
Version 2.1.0  (2005-04-18), ISBN 3-900051-07-0

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for a HTML browser interface to help.
Type 'q()' to quit R.

[Previously saved workspace restored]

> 
```

El signo *mayor que* en la última línea indica que podemos escribir una expresión con la sintaxis de R. Tras pulsar retorno de carro este la evaluará y actuará en consecuencia.

Para abandonar el programa bastará con escribir `q()` y el resultado será:

```
> q()
Save workspace image? [y/n/c]:
```

que nos permite:

y guardar los datos incorporados al programa y el histórico con las ordenes utilizadas durante la sesión de trabajo.

n perder datos e histórico de ordenes.

c cancelar la orden dada para abandonar el programa y permanecer en la sesión de trabajo.

La ventaja de conservar los datos y el histórico es que permite retomar la sesión de trabajo con posterioridad en el punto donde la abandonamos.

Al iniciar una sesión, si existe un histórico bastará pulsar la tecla $\boxed{\uparrow}$ para acceder las expresiones escritas con anterioridad y la tecla $\boxed{\downarrow}$ para las escritas posteriormente. Estas expresiones recuperadas pueden además, editarse utilizando las teclas $\boxed{\leftarrow}$ y $\boxed{\rightarrow}$.

2.2.1 Ayuda en R

Una de las cualidades de R es su sistema de documentación y ayuda. Para obtener ayuda sobre distintos aspectos del programa puede utilizarse la orden `help()` o `help.start()`. Además se dispone de algunas demostraciones del uso, para localizarla se utiliza la orden `demo()`. También es posible utilizar los ejemplo que aparecen en la ayuda con la función `example()`

2.2.2 Expresiones en R

R utiliza una sintaxis muy sencilla en la que la unidad está constituida por la *expresión*. La expresión más sencilla es un simple número entero.

```
> 1
[1] 1
```

R evalúa la expresión y devuelve un valor de esta. La notación `[1]` indica que el primer valor de la línea de respuesta, es el primer valor del conjunto de números que constituyen la respuesta de R.

Si tecleas

```
> seq(1,100)
  [1]   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18
 [19]  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36
 [37]  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54
 [55]  55  56  57  58  59  60  61  62  63  64  65  66  67  68  69  70  71  72
 [73]  73  74  75  76  77  78  79  80  81  82  83  84  85  86  87  88  89  90
 [91]  91  92  93  94  95  96  97  98  99 100
```

comprobarás que las respuestas de R pueden llegar a ser muy largas y que por tanto resulta útil tener un indicador del índice (o número de orden) de los elementos que constituyen la respuesta.

Los valores pueden relacionarse mediante operadores (lógicos o algebraicos) y funciones.

```
> sqrt(3^2 + 5^2)
[1] 5.830952
```

Insistiremos más adelante en las funciones (aquí se ha utilizado la correspondiente a la raíz cuadrada).

En R tenemos los siguientes operadores que puede utilizarse para:

`+, -, *, /`	suma, resta, producto, cociente
`%%, %/%, ^`	módulo, cociente entero, potencia
`==, !=, !`	igual, distinto, no
`>, >=, <, <=`	mayor que, mayor o igual que, menor que, menor o igual que
`|, &`	o, y

Para obtener información de los distintos operadores puede solicitarse la ayuda mediante `?"+"`.

Cada operador implica a los dos elementos que se sitúan a izquierda y derecha en la expresión, existe una jerarquía de operadores que determina cual de ellos se evalúa primero en la expresión, el orden de evaluación puede alterarse mediante el uso de paréntesis. Por ejemplo:

```
> 9 * 5 / 2
[1] 22.5
> 9 * 5 / 2 * 3
[1] 67.5
> 9 * 5 / (2 * 3)
[1] 7.5
```

que efectúan $\frac{9 \times 5}{2}$, $\frac{9 \times 5 \times 3}{2}$, $\frac{9 \times 5}{2 \times 3}$

Los operadores lógicos proporcionan como resultado el valor VERDADERO (anotado por `TRUE`, o `T`) o el valor FALSO (anotado por `FALSE`, o `F`)

```
> 3 != 2
[1] TRUE
```

Cuando se necesita más de una expresión estas se pueden escribir en líneas separadas, si bien, puede reunirse en una sola línea más de una expresión separándolas por ";".

En algunos casos es posible escribir una expresión en más de una línea, así cuando R entiende que la expresión es incompleta devuelve un *prompt* $+$ en lugar del habitual $>$, tras este signo más podemos continuar la expresión que se quedó a medias en la linea precedente.

Ejercicios:

1. Calcular el resultado de elevar a 5 el producto de 6 por 9.
2. Haz diversos ejemplos que te permitan entender bien la diferencia entre `/`, `%%` y `%|%`.
3. Verificar si un número 3895 es múltiplo de 3.
4. Escribe la expresión para verificar que 5 es igual a 2+3 y aplicarla.

2.2.3 Variables

Podemos utilizar variables para guardar el resultado de una expresión:

```
> a = 1
```

El programa no devuelve mensaje alguno, pero desde ahora la variable `a` contiene valor 1, así si escribimos:

```
> a
[1] 1
```

el sistema nos devuelve el valor de la variable.

La asignación significa una actualización del valor de la variable, tras efectuarla se pierde el viejo valor y se queda el que resulta de la expresión asignada.

```
> a = 1
> a = 5
> a
[1] 5
```

Los nombres de las variables deben comenzar obligatoriamente por una letra, distinguiéndose entre mayúsculas y minúsculas, y a continuación, opcionalmente, una combinación de letras y números, a ellos puede incorporarse el punto "."; así son nombre válidos: `a, A, A.1, altura, densidad,` ... Los nombre de las variables pueden coincidir con los nombre de funciones, aunque resulta poco aconsejable ya que pueden producirse conflictos.

Resulta conveniente que los nombres de las variables mantenga una relación con los valores que contienen: así, la variable que describe la 'cobertura lineal del palmito (*Chamaerops humilis*)' puede indicarse como: `cl`, haciendo referencia a la idea de cobertura lineal, si no tenemos otras especies; `cl.chahum`, como en el caso anterior pero cuando tenemos varias especies; o `chahum`, cuando solo tenemos información para esta especie de su cobertural lineal.

Además de valores numéricos podemos utilizar las cadenas de texto. Para indicarle al ordenador que toda la cadena es un sólo objeto, esta debe entrecomillarse :

```
> a="Hola"
> nombre.estacion="Pluviómetro de San Javier"
```

Las variables pueden contener valores sencillos, atendiendo a su naturaleza tenemos:

Lógicos:	TRUE,FALSE,T,F
Enteros:	-10, 1, 1000, ...
Reales:	-10.1, 6.02310e24, ..., -Inf, Inf, NaN
Complejos:	1+3i, 1+0i, 9i, ...
Carácter:	"Hola", "Enero", "sin(x)", "pino", ...
Desconocido:	Na

Ejercicios

1. ¿Qué resultado se obtendrá, si asignamos el valor 5 a la variable `a` y posteriomente 3 a la misma variable, y calculamos `2 * a`?
2. Siguiendo el caso anterior, ¿cuál es el resultado de realizar la siguiente asignación?

   ```
   a = a + 1
   ```

3. ¿Cuáles de las siguientes asignaciones son legales?:

   ```
   b = 2*a      a-5 = -3*b      a = 4 = b      a = b = 6      a = ((b<-5)*2)
   ```

2.2.4 Vectores

En R podemos agrupar variables sencillas para construir objetos más complejos:

Para este capítulo consideraremos,por simplificar, sólo vectores, matrices y data frames.

Para disponer de un vector basta con asignar a una variable un conjunto de valores:

```
> x = c(1,22,9,8,36,12)
```

crea el vector `x` siendo $x_1 = 1, x_2 = 22, \ldots, x_6 = 12$, con la ayuda de la función `c()` que permite concatenar una serie de valores. Los valores de `x` pueden consultarse sencillamente:

```
> x
[1]  1 22  9  8 36 12
```

Podemos también obtener cualquiera de los elementos del vector de forma independiente mediante un subíndice:

```
> x[3]
[1] 9
```

como puede apreciarse claramente, los subíndices se indican entre corchetes.

Un operador interesante es ":", así, si escribimos:

```
> 1:5
[1] 1 2 3 4 5
```

obtenemos cinco valores consecutivos: los valores del uno al cinco. Podemos utilizar este operador para generar series. Los rangos así obtenidos pueden utilizarse en distintos casos, por ejemplo como subíndices:

```
> x[1:3]
[1]  1 22  9
```

con esta expresión seleccionamos los tres primeros elementos del vector `x`.

Cuando no se expresa subíndice alguno nos referimos al vector completo, por ello `x` y `x[ ]` son expresiones equivalentes.

Un subíndice negativo equivale a eliminar el elemento indicado del vector, así: `x[-5]`, indica el vector `x` excluido el quinto elemento.

```
> x[-5]
[1]  1 22  9  8  36
```

En ocasiones necesitamos seleccionar de un vector elemento condicionados a una propiedad, por ejemplo, si deseamos utilizar sólo los elementos del vector con valor par lo expresamos: `x[x%%2==0]`.

Algunos vectores particulares pueden obtenerse de forma directa con las funciones `rep` y `seq`.

```
x=rep(1,5)
x
[1]  1  1  1  1  1
y=seq(1,9,by=2)
y
[1]  1  3  5  7  9
```

Queda claro que `rep(a,n)` crea un vector con n elementos todos iguales a a; y que `seq(a,b,by=i)` genera la secuencia de números enteros de a a b de i en i.

Ejercicios

1. Obtener la serie de números de 25 a 35.
2. Utilizar las funciones `rep()` y `seq` para obtener una secuencia 1, 2, 3, 1, 2, 3, 1, 2, 3, es decir tres secuencias del 1 al 3 consecutivas.
3. ¿Cúal es resultado de `1:5 > 2` ? ¿por qué?
4. Utilizar la función `rep()` para obtener una secuencia 1, 2, 3, . . . , 10 repetida 10 veces.
5. Determinar qué números de la serie 3895 a 3910 son múltiplos de 3.
6. ¿Cuál es el resultado de calcular el logartimo natural mediante la función `log()` al vector (6, 0, 2.8, -1, 5.59)?
7. ¿Cuál será el resultado de multiplicar dos vectores del mismo tamaño? ¿y si el tamaño varía?

2.2.5 Funciones

En R la mayor parte del trabajo implica el uso de funciones, por ejemplo:

```
> x=c(1,22,9,8,36,12)
> max(x)
[1] 36
```

nos devuelve el máximo valor que presenta un elemento del vector *x*.

El número y utilidad de las funciones de R es enorme y variado. Además, y esta es una de las virtudes de R, pueden definirse fácilmente nuevas funciones.

Las funciones se indican con una palabra clave, o nombre de la función, y entre paréntesis —opcionalmente— parámetros. Por ejemplo, la función `q()`, que se utiliza para abandonar R, no necesita parámetro alguno; la función `sqrt()` necesita un valor para obtener su raíz cuadrada —está claro que este valor puede ser el resultado de una expresión compleja.

Entre otras podemos destacar las siguientes funciones para números:

`sqrt()`	Raíz cuadrada.
`abs()`	Valor absoluto.
`sin(), cos(), ...`	Funciones trigonométricas (el parámetro debe expresarse en radianes).
`log(), exp()`	Logaritmo y exponencial.
`round()`	Redondeo de valores numéricos.

Funciones para vectores:

`length(x)`	Devuelve el número de elementos del vector x
`sum(x)`	Devuelve la suma de todos los elementos del vector x
`prod(x)`	Devuelve el producto de todos los elementos del vector x
`max(x)`	Devuelve el máximo de los elementos del vector x
`min(x)`	Devuelve el mínimo de los elementos del vector x
`which.max(x)`	Devuelve la posición en que se encuentra el máximo de los elementos del vector x
`which.min(x)`	Devuelve la posición en que se encuentra el mínimo de los elementos del vector x
`rev(x)`	Devuelve el vector en orden inverso
`sort(x)`	Devuelve el vector ordenado de menor a mayor
`cumsum(x)`	Devuelve un vector con los valores de x acumulados
`table(x)`	Devuelve una tabla con cada uno de los valores diferentes de x y el número de veces que aparece

Ejercicios

1. Calcular con R:

$$\sqrt{log(3) - \frac{387}{23}} \tag{2.1}$$

2. Aplica la función:

$$e^{\sqrt{x}} \tag{2.2}$$

   a todos los valores enteros entre 1 y 100

3. Obten con R el vector [3,6,4,9,2] ordenado de forma descendente

2.2.6 Matrices y data.frames

Una matriz es una estructura rectangular de números que tiene numerosas aplicaciones en matemáticas. Podemos generar una matriz con la función `matrix(v,filas,columnas)` donde v es un vector de números cuyo tamaño debe ser igual al número de filas por el de columnas. Por ejemplo:

```
> v=c(3,4,3,5,6,5)
> matrix(v,3,2)
     [,1] [,2]
[1,]    3    5
[2,]    4    6
[3,]    3    5
```

Lógicamente el número de elementos del vector que se introduce a la función $matrix$ debse ser igual al producto del número de filas por el número de columnas que se especifiquen.

Entre las funciones para manejar matrices cabe destacar:

`rbind(m1,m2)`	Crea una matriz situando m2 debajo de m1
`cbind(m1,m2)`	Crea una matriz situando m2 a la derecha de m1
`t()`	Devuleve la matriz traspuesta

Los data frames son matrices en las que cada columa representa una variable cuyo nombre puede incluirse en el objeto. Podemos acceder a cada variable por separado con la expresión `data.frame$variable`.

Por ejemplo si el data frame `lluvia` contiene doce variables, una por cada mes podremos acceder a los datos de febrero escribiendo `lluvia$febrero`.

Puedes crear un data frame agrupando varios vectores del mismo tamaño con la función `data.frame()`:

```
> x=c(1,2,3,4,5)
> y=c(5,4,3,2,1)
> d=data.frame(x,y)
> d$y
[1] 5 4 3 2 1
```

Para leer datos de un fichero o escribir datos de un fichero, generalmente se utilizan las funciones `read.table` y `write.table`. La primera lee un data.frame de un fichero de texto y la segunda escribe un data.frame en un fichero de texto. Estas funciones se verán más adelante.

Ejercicios

1. Trata de definir matriz traspuesta a partir del resultado de trasponer algunas matrices previamente definidas por ti.

2. Genera un data frame que contenga tres variables: los nombres de los meses, el número de días de cada mes y el número de días acumulado.

2.2.7 Funciones genéricas

Algunas funciones de R son de propósito general y se utilizan con cualquier tipo de objeto (vectores, matrices, data frames u otro tipo de objetos más complejos). Dependiendo de cual sea el objeto al que se apliquen los resultados son diferentes.

Entre estas funciones, las más relevantes son:

- `print`, vuelca el contenido del objeto en pantalla, es en muchos casos equivalente a escribir el nombre del objeto sin más.

- `plot`, hace un gráfico del objeto, dependiendo del tipo de objeto la representación gráfica varía. La aplicación de esta función a vectores se estudiara con más detalle posteriormente.
- `summary`, proporciona estadísticos básicos de un objeto. Resulta útil por ejemplo para conocer el contenido de un data frame recien importado.
- `str`, desvela la estructura de un objeto, que otros objetos lo forman.

2.2.8 Funciones de usuario

A pesar del amplísimo número de funciones de que dispone R, en ocasiones necesitaremos programar nuestras propias funciones. En R es bastante sencillo, en los temas siguientes se verán algunos ejemplos de funciones creadas para esta asignatura y que se encuentran en el fichero `funciones.R` que puedes descargar de la página web de la asignatura. Puedes ver el contenido del fichero con cualquier editor de textos (el block de notas de windows bastará). En el tema **??** se verá como crear funciones.

2.3 Importación y exportación de datos

Una de las funciones básicas de un programa de análisis de datos debe ser, lógicamente, la lectura de datos a partir de un archivo de texto. Para leer un fichero simple, con los datos separados por espacios en blanco, tabuladores o saltos de línea, se utiliza la instrucción `read.table` en la forma:

```
> datos = read.table("base_datos.txt")
```

si el fichero tiene una linea de cabecera con los nombres de las variables utilizaremos:

```
> datos read.table("base_datos.txt", header = TRUE)
```

y de esta manera se obtienen, de la primera linea, los nombres de las variables.

Si el separador de variables no es un espacio y es, por ejemplo un "`;`" hay que utilizar el parámetro `sep=";"`. Si el carácter decimal no es un punto sino, por ejemplo, una coma, usar `dec = ","`.

Si se quiere guardar una matriz o vector en un fichero de texto puede utilizarse la función write.table:

```
> write.table(resultados,"salida.txt")
```

Esta función admite los parámetros `sep` y `dec` y además los parámetros lógicos `append`, `row.names` y `col.names`; puesto que son parámetros lógicos pueden tomar los valores `T` (cierto) o `F` (falso). El primero determina que ocurrirá si el fichero ya existe, si `append=T` la tabla se añadirá al fichero en caso contrario el fichero se sobreescribirá; los otros dos parámetros seterminan si la salida debe incluir los nombres de las filas y las columnas, respectivamente.

El fichero base_datos.txt contiene algunas series de precipitación con las que se harán parte de los ejercicios de este curso. Las columns contienen los diguientes datos:

ano Año

precZ Precipitación total en Zarzadilla de Totana

maxZ Precipitación máxima en Zarzadilla de Totana

diasZ Número de días de precipitación en Zarzadilla de Totana

precL Precipitación total en Librilla

maxL Precipitación máxima en Librilla

diasL Número de días de precipitación en Librilla

precY Precipitación total en Yecla

maxY Precipitación máxima en Yecla

diasY Número de días de precipitación en Yecla

precB Precipitación total en Beniaján

maxB Precipitación máxima en Beniaján

diasB Número de días de precipitación en Beniaján

precE Precipitación total en El Algar

maxE Precipitación máxima en El Algar

diasE Número de días de precipitación en El Algar

Ejercicios

1. Crea un data frame a partir del fichero `base_datos.txt` que puedes descargar de la página web de la asignatura.
2. Explora su contenido con las cuatro funciones genéricas que se vieron anteriormente.

2.4 Gráficos con R

Una de las mayores potencialidades de **R** es su capacidad para producir gráficos de diverso tipo (probad el comando `> demo(graphics)`) para obtener una demostración de sus capacidades gráficas.

De momento nos quedamos con dos comandos:

- `> hist (x)`, para producir histogramas (ver figura 2.2).
- `> plot (x,y)`, para representar una variable contra otra.

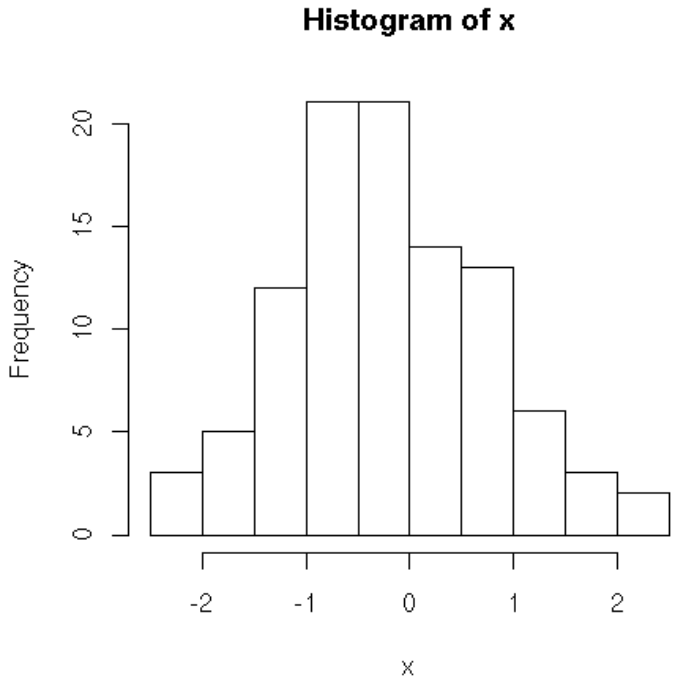


Figura 2.2: Histograma producido con la función `hist`

Estos comandos tienen multiples opciones, por ejemplo

```
> x=rnorm(100)
> hist (x,breaks=10)
```

hará el histograma con aproximadamente, 10 barras,

```
> hist (x,freq=F)
```

representará la densidad (proporción en tantos por uno) en lugar de la frecuencia.

Puede tambien modificarse el color de las barras con la opción color:

```
> hist (x,col=”red”)
```

o incluir nuestro propio título al histograma:

```
> hist (x,main=”Histograma de x2)
```

Todas estas opcciones pueden combinarse.

Pueden también modificarse los rótulos de los ejes de abcisas y ordenadas, así como añadir un título al gráfico

```
> x=rnorm(100)
> hist (x,xlab=”Rótulo del eje de abcisas”,ylab=”Rótulo del eje de ordenadas”,
main=”Título del gráfico”)
```

En cuanto al comando plot, permite la mayor parte de las opciones anteriores y algunas. Por ejemplo la siguiente linea:

```
plot(x, y, main = "Título principal", sub = "subtítulo", xlab = "eje x",
ylab = "eje y", xlim = c(-5,5),ylim = c (-5,5))
```

produce como resultado el gráfico de la figura 2.4.

La orden `legend` permite incluir una leyenda en el gráfico:

```
> legend(1, 4, legend=c("uno", "dos", "tres"), lty = 1:3, col = c("red",
"blue", "green"),pch = 15:17)
```

Los dos primeros parámetros indican la posición que tendrá la leyenda dentro del gráfico, parámetro `legend` hace referencia a los textos de la leyenda, el parámetro `lty` a los tipos de linea, `col` a los colores en que se va a pintar y `pch` a los tipos de símbolos (figura fig:legendR).

Con `text` podemos representar caracteres de texto directamente. Como ejemplo pruebese el siguiente programa (Para R las lineas que comienan con # se consideran comentarios y R no tratan de ejecutarse):

```
# Obtener 50 pares de números procedentes de una distribución uniforme
# con valores entre 0 y 4
x = runif(50, 0, 4); y = runif(50, 0, 4)

# Obtiene un vector de 50 valores, los 20 primeros son "v"
# y los siguientes "m"
```

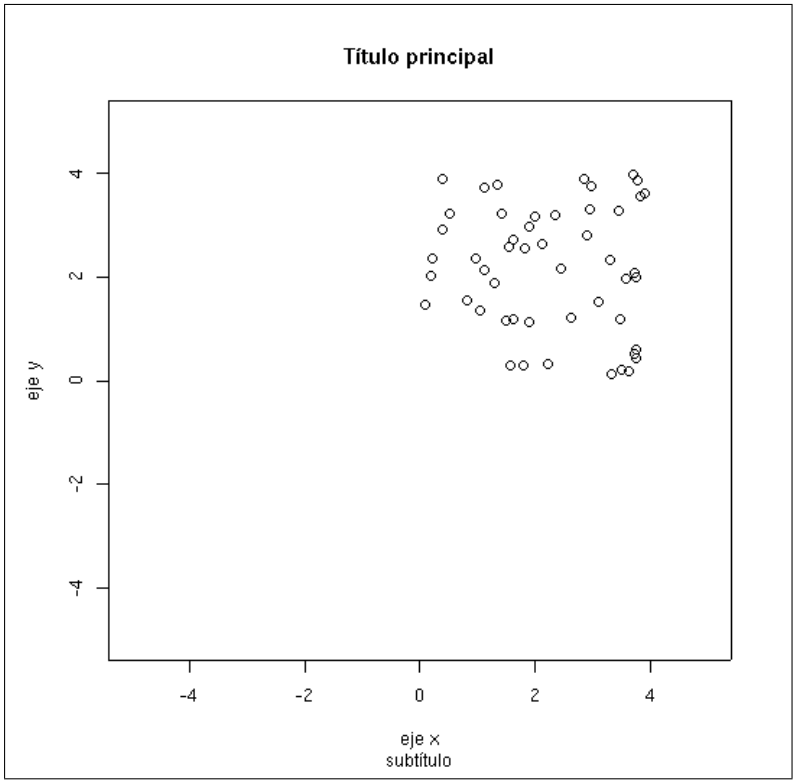


Figura 2.3: Ejemplo de gráfico extraido con plot

```
sexo = c(rep("v", 20), rep("m", 30))

# Crea los ejes pero no pinta el gráfico
plot(x, y, type = "n")

# Escribirá una "v" o un "m" en las posiciones indicadas por
# las variables x e y.
text(x, y, labels = sexo)
```

que producirá una figura similar a la de la figura 2.5 (puesto que los valores de x e y son aleatorios variarán de una ejecución a otra).

El comando `plot` introduce muchas opciones. Por ejemplo el parámetro `type` permite 4 posibilidades: `p` para puntos (opción por defecto), `l` para lineas, `b` para lineas y puntos, `h` para barras verticales y `n` para representar sólo los ejes pero no los datos.

```
> plot (x,y,type=”p”)
```

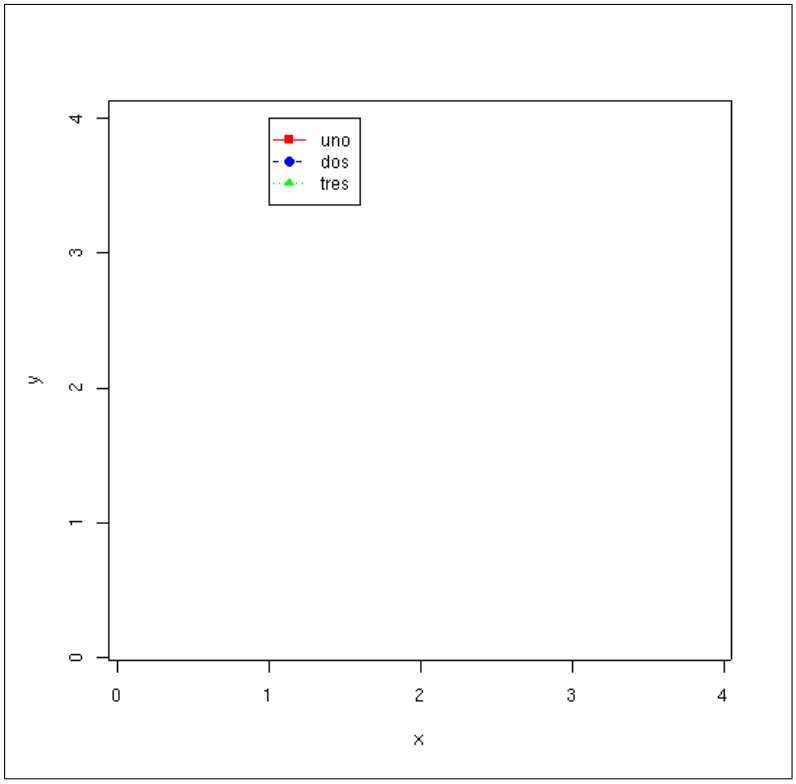


Figura 2.4: Ejemplo de leyenda sobre un gráfico extraido con plot

```
> plot (x,y,type=”l”)
> plot(x, y, type = "b")
> plot(x, y, type = "h")
> plot(x, y, type = "n")
```

2.4.1 Modificación del aspecto de puntos y lineas

Si la representación se hace por puntos, puede resultar de interés la modificación del tipo de símbolos, por ejemplo para pintar el carácter k en lugar de puntos:

```
> plot (x,y,pch=”k”)
```

Existen 25 símbolos predefinidos (figura 2.6) a los que se accede mediante la variable `pch` y 7 colores predefinidos accesibles con la variable `col`. Por ejemplo:

```
> points(x, y, pch = 2, col = 3)
```

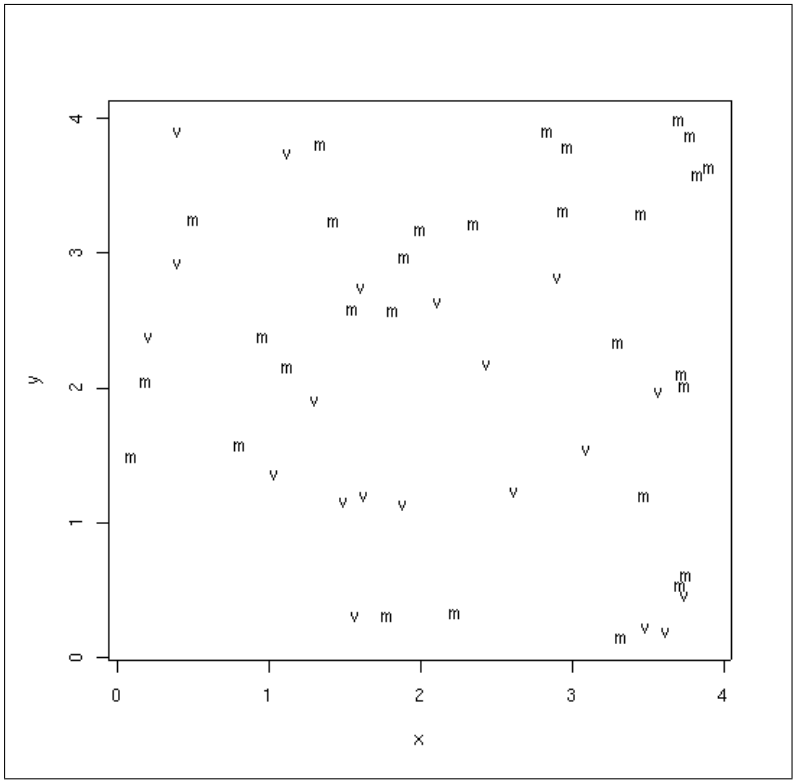


Figura 2.5: Ejemplo de textos sobre un gráfico extraido con plot

dibujará triángulos sin rellenar de color verde. Tal como puede comprobarse en la figura 2.6, si se indica8 como color, el sistema asigna el 1 y así sucesivamente

Puesto que colores y símbolos admiten valores numéricos, podemos asignarles variables, de manera que cada punto puede tener un símbolo y color diferentes en función de sus valores para esas variables.

```
> plot (x,y,pch=z,col=suelo).
```

Tal como puede verse en la figura 2.7, los parámetros `lty` y `lwd` permiten modificar los tipos y anchos de linea respectivamente.

2.4.2 Guardar los gráficos

El modo de guardar un gráfico, para por ejemplo introducirlo en un documento de texto o en una presentación, depende de que se esté utilizando la versión de R de windows o de linux.

- **En linux**

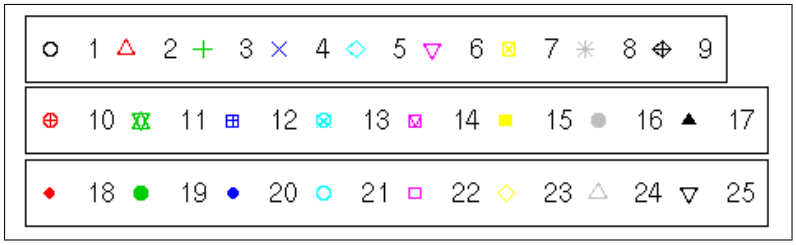


Figura 2.6: Símbolos y colores en los gráficos con R

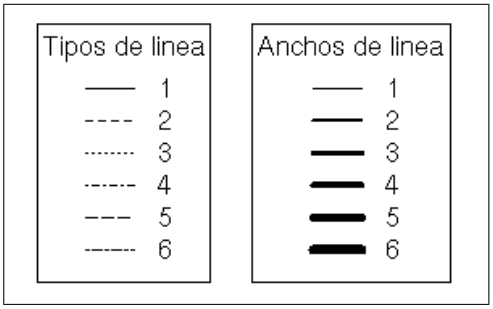


Figura 2.7: Tipos y anchos de linea en los gráficos con R

Si lo que se quiere guardar es un gráfico, por ejemplo en formato jpeg, debe definirse en primer lugar que la salida del gráfico va a ser un archivo de tipo jpeg y su nombre:

```
>jpeg("grafico.jpg")
```

posteriormente realizar el gráfico

```
> plot(a,z)
```

y finalmente *desconectar* el comando jpeg.

```
> dev.off()
```

- **En windows**

  En windows la ventana que muestra el gráfico de R incluye un menú en el que seleccionaremos la opción `Archivo -> Guardar Como` que nos permitirá elegir el formato de fichero y finalmente el nombre del archivo gráfico.

Ejercicios

A partir del data frame que creaste en el ejercicio anterior, realiza los siguientes gráficos:

1. Haz un gráfico con la serie de precipitación de Yecla

2. Haz un histograma con las series de precipitación y precipitación máxima de Yecla. Prueba diferentes opciones

3. Haz un gŕafico con las cinco series de precipitación juntas de manera que se distingan de alguna manera unas de otras. Introduce una leyenda

4. Representa gráficamente la precipitación de Beniaján contra la de Zarzadilla de Totana. Comenta el gráfico resultante

Tema 3

Estadística descriptiva. Análisis de muestras

3.1 Introducción

La ciencia estadística ha desarrollado un amplio conjunto de métodos y técnicas para el análisis de datos recogidos en cualquier campo de investigación científica.

Estos métodos son especialmente útiles en ciencias cuyo objeto de estudio se caracterice por comportamientos *complejos* a los que difícilmente se pueden aplicar leyes generales como las leyes físicas. Este tipo de ciencias incluyen ciencias ambientales (Geografía, Ecología, Geología) o sociales (Geografía, Sociología, Psicología).

Suele dividirse la estadística en una serie de ramas, la división clásica es estre **estadística descriptiva e inferencial**; otra división clasica se establece entre **estadística univariante** (cuando se estudia el comportamiento de una sola variable) y **estadística multivariante** (cuando se estudia el comportamiento de varias variables así como sus interacciones). Más recientemente, gracias a la incoroporación de los ordenadores, se ha desarrollado la **simulación estadística**.

Conceptos previos:

- **Individuo**. Unidad mínima de información estadística
- **Población**. Conjunto de individuos con una característica común
- **Muestra**. Subconjunto de una población en la que se va a realizar un experimento
- **Muestreo**. Proceso de extracción de una muestra
- **Variable**. Propiedad medible u observable en un individuo
- **Experimento**. Medición de una o varias variables en un individuo o muestra de individuos
- **Suceso**. Se obtiene un determinado valor o valores en un conjunto de variables medidas sobre una muestra.

- Estadística **descriptiva**: Caracterizar una **muestra** mediante un conjunto de **estadísticos**. Se trata de valores calculado a partir de una o varias variables medidas en los individuos de la muestra.

- Estadística **inferencial**: Estimar propiedades de la **población** de la que procede la muestra, estas propiedades son un conjunto de **parámetros** que se corresponden con los estadísticos muestrales. Los estadísticos se suelen representar mediante letras latinas y los parámetros mediante letras griegas. Así la media poblacional se representa con μ y la media muestral con m aunque también puede representarse la media de la variable x como $\hat{x}$.

- **Contraste de hipótesis**: Determinar, a partir de los datos procedentes de un muestreo, si puede aceptarse una determinada hipótesis.

- **Simulación** estadística: Utilizar un modelo estadístico, basado en las características poblacionales inferidas, para generar muestras artificiales.

En este tema se presentarán algunos de los estadísticos utilizados en estadística descriptiva distinguiendo aquellos que se aplican a una sola variable y los que se utilizan para estudiar la relación entre dos variables.

3.2 Estadística univariante

3.2.1 Gráficos

Histograma

Es la herramienta más simple y utiñizada de representación de datos univariantes. Tiene el inconveniente de obligarnos a elegir un número de clases (3.2).

Gráficos de probabilidad acumulada

Una alternativa son los gráficos de frecuencia acumulada, que se obtiene ordenando los datos y acumulando las frecuencias de los datos ordenados, y el gráfico de probabilidad acumulada, que se obtiene dividiendo las frecuencias acmuladas entre $N + 1$ donde N es el número total de datos.

En R puedes obtener la probabilidad acumulada de una variable con la función `prob_acu` del fichero **funciones.R**.

3.2.2 Estadísticos

- Histograma, es la primera aproximación gráfica y descriptiva al comportamiento de una variable

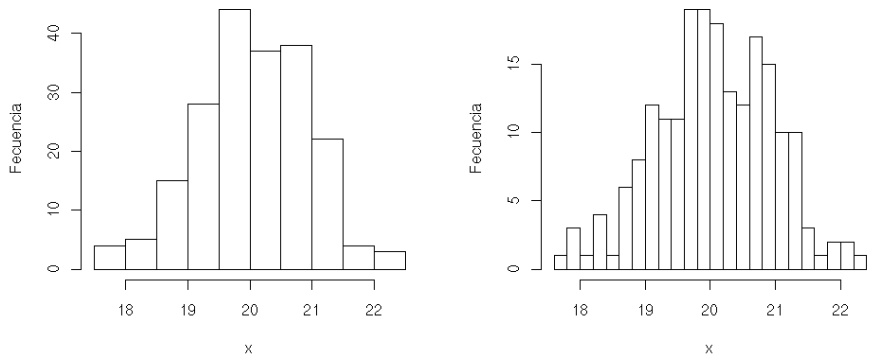


Figura 3.1: Histogramas

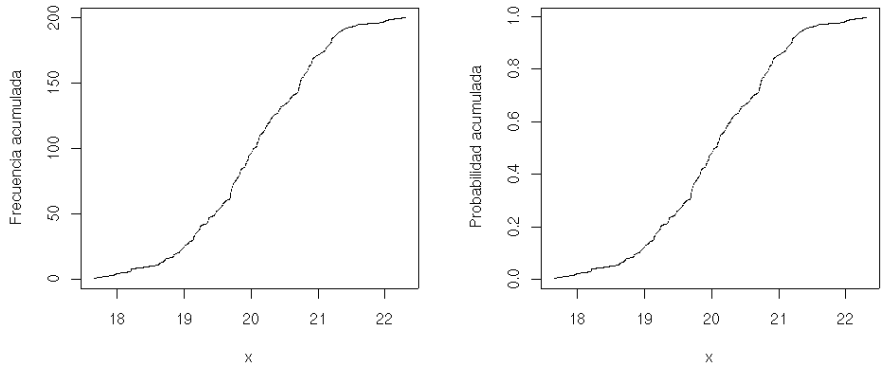


Figura 3.2: Gráficos de frecuencia acunulada y probabilidad acumulada

- Estadísticos de centralidad: media (m), mediana q_{50} (valor intermedio) y moda (valor más frecuente)

$$m_x = \frac{\sum x_i}{n} \tag{3.1}$$

Los estadísticos de centralidad nos indican en que punto del eje de abcisas se sitúa el histograma de la muestra (figura 3.3).

- Estadísticos de dispersión: varianza (s^2), desviación típica (s), rango intercuartílico ($q_{75} - q_{25}$), coeficiente de variación ($CV = s/m$)

$$s_x = \sqrt{\frac{\sum (x_i - m_x)^2}{n}} \tag{3.2}$$

Cuando la muestra es muy pequeña (n<30) suele utilizarse la siguiente ecuación modificada para el cálculo de la desviación típica:

$$s_x^2 = \frac{\sum (x_i - m_x)^2}{n-1} \tag{3.3}$$

$$s_x = \sqrt{\frac{\sum (x_i - m_x)^2}{n-1}} \tag{3.4}$$

aunque en realidad muchos programas de análisis estadístico utilizan esta última aproximación para cualquier tamaño muestral.

Los estadísticos de dispersión nos indican el grado de dispersión respecto a la media que muestran los valores de la variable (figura 3.4).

Uno de los problemas de la desviación típica es que no es comparable entre dos variables si estas presentan magnitudes muy diferentes. Para evitar este inconveniente se utiliza el coeficiente de variación (CV):

$$CV_x = 100\frac{s_x}{m_x} \tag{3.5}$$

Este estadístico es ya independiente de la magnitud de la variable y permite por tanto comparar el comportamiento de diferentes variables.

- Estadísticos de forma: coeficientes de sesgo (g), que mide el grado de asimetría de la muestra, y de curtosis (k), que mide el grado de apuntamiento o aplanamiento del histograma de la muestra e indirectamentemente el grado de homogeneidad o heterogeneidad de la muestra.

$$g_x = \frac{\sum (x_i - m_x)^3}{N s_x^3} \tag{3.6}$$

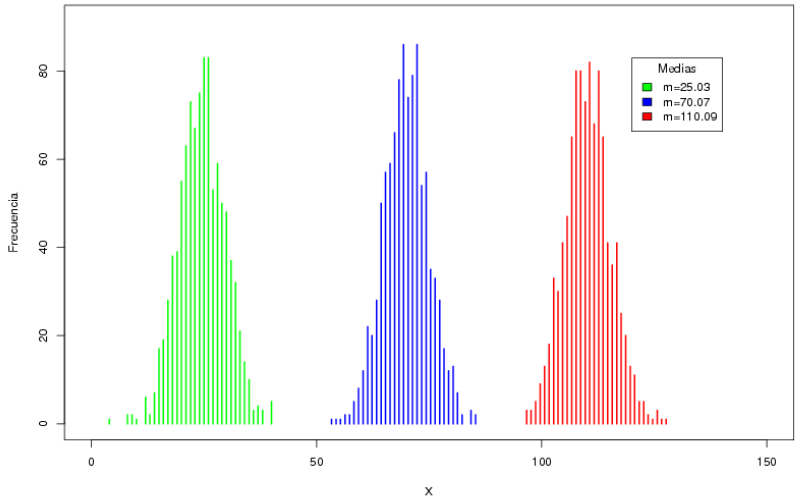


Figura 3.3: La media muestra donde se sitúa en el eje X el centro de la muestra

Si $g = 0$ el histograma tiene forma simétrica, si $g > 0$ la mayor parte de los datos son menores que la media, si $g < 0$ la mayor parte de los datos son mayores que la media (figura 3.5).

$$k_x = \frac{\sum (x_i - m_x)^4}{N s_x^4} \tag{3.7}$$

Un valor de $k = 3$ se considera propio de una **distribución normal** (ver figura 5.2). Un valor de $k > 3$ implica una distribución (histograma) aplanado (muestra heterogenea) en cuyo caso caba suponer que algunos valores extremos, posiblemente no pertenecientes a la misma población, se han colado en la muestra.

Si $k < 3$ el histograma tiene forma apuntada, lo que implica que la muestra puede ser muy homogenea. Sin embargo también podemos obtener un valor menor que 3 si se han mezclado dos poblaciones de características muy diferentes y que apareceran perfectamente definidas al estudiar el histograma (en estos casos k es incluso menor que dos) (figura 3.6).

Los estadísticos muestrales pueden ser **paramétricos** (aquellos basados de una u otra forma en el cálculo de la media) o **no paramétricos** (como por ejemplo q_{25}, q_{50} y q_{75}). Los primeros son más adecuados cuando la variable es *normal* y los segundos en caso contrario. Por el momento vamos a quedarnos con la idea de que una variable es normal cuando presenta un histograma simétrico con forma de campana (figura 5.2).

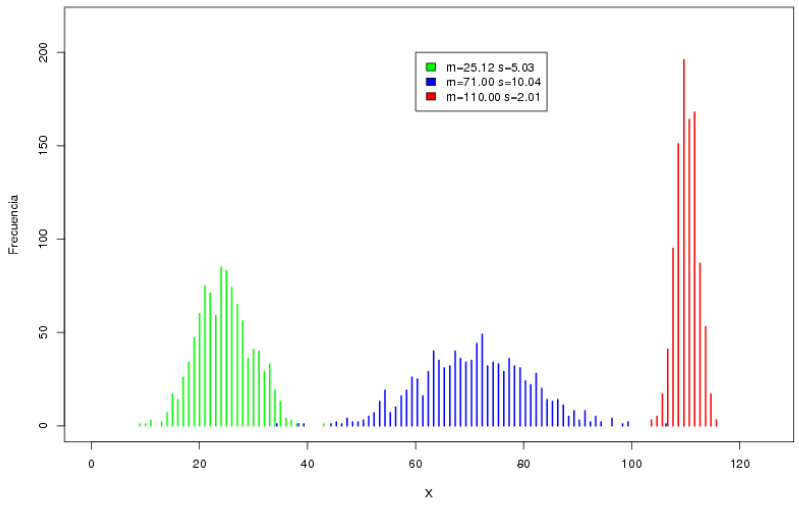


Figura 3.4: La desviación típica indica el grado de dispersión respecto al valor de la media

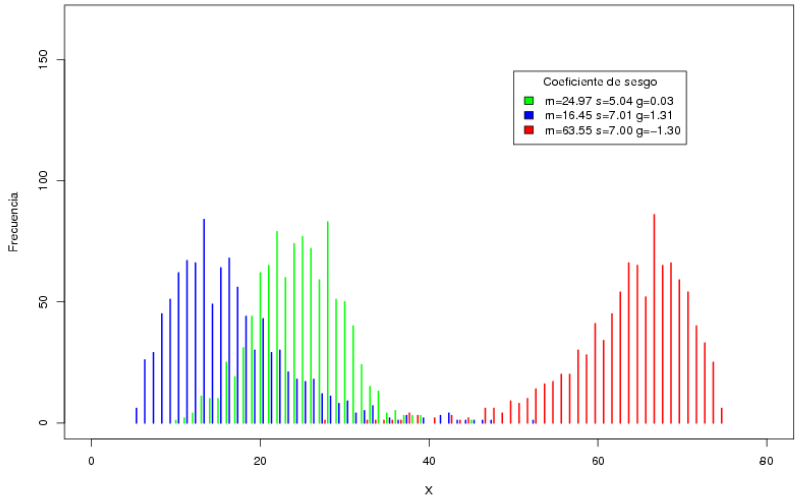


Figura 3.5: El coeficiente de sesgo indica la asimetría de la muestra

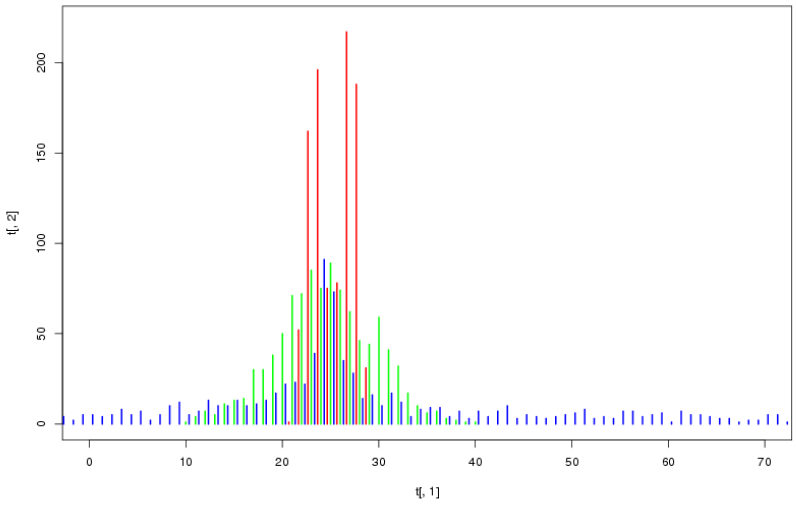


Figura 3.6: El coeficiente de kurtosis indica el grado de aplanamiento o apuntamiento de la muestra

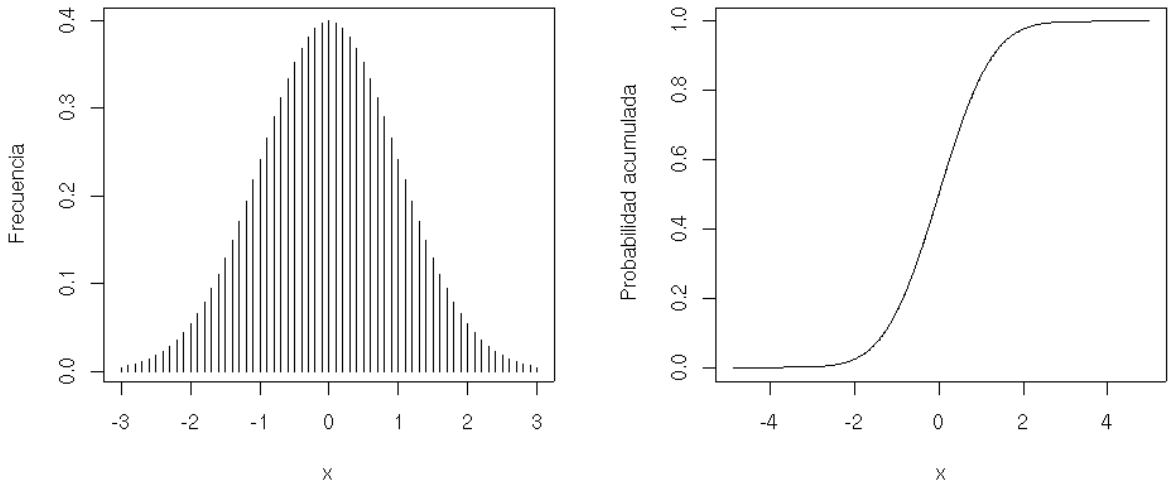


Figura 3.7: Distribución normal o campana de Gauss

3.2.3 Con R

R dispone de múltiples funciones para el análisis estadístico. La representación de histogramas la puedes repasar en el tema anterior. Entre las funciones apropiadas para estadística descriptiva aparecen:

- `mean(x)` para calcular la media
- `var(x)` para calcular la varianza
- `sd(x)` para calcular la desviación típica
- `sesgo(x)` para calcular el coeficiente de sesgo
- `curtosis(x)` para calcular el coeficiente de curtosis
- `quantile(x,p)` para calcular el cuantil p, por ejemplo para la mediana (1_{50}): `quantile(x,0.5)` y para el cuantil 0.25 (q_{25}): `quantile(x,0.25)`

Las funciones `sesgo` y `curtosis` no aparecen en R y debes obtenerlas cargando el fichero **funciones.R** que puedes descargar de la página web de la asignatura. Para cargar el código de R contenido en este fichero deberás utilizar la función `source`:

```
> source("funciones.R")
```

Ten en cuenta que tanto en linux como en winxx deberás incluir junta al nombre del fichero la ruta de directorios completa.

3.2.4 Ejercicios

- Calcula los estadísticos básicos de la precipitación anual obtenida en 2 observatorios y que se incluye en el fichero ejercicio_3_2_2_a.txt que puedes descargar de la página web.
- Representa gráficamente ambas variables
- ¿Cual es la diferencia entre media y mediana en ambos casos?
- ¿A que observatorio consideras más probable que corresponda una precipitación de 400 mm? Justifica tu respuesta.
- Crea un data frame con los datos de precipitación anual obtenida en 2 observatorios y que se incluye en el fichero ejercicio_3_2_2_b.txt que puedes descargar de la página web.
- Calcula el coeficiente de curtosis de los dos observatorios y de una serie generada uniendo ambos observatorios.
- ¿Qué conclusiones sacas a partir de los coeficientes calculados?

3.3 Estadística bivariante

3.3.1 Gráficos bivariantes

El gráfico de dispersión (figura 3.10) muestra la relación entre dos variables cuantitativas, mientras que los gráficos de caja y bigotes (*box and whiskers* (figura 3.11) muestran la relación entre una variable cualitativa y una variable cuantitativa.

En el gráfico de caja, se representa la distribución de una variable cuantitativa. La parte superior de la caja representa el cuantil 75, la inferior el cuantil 25 y la linea central de la caja la mediana. Las lineas superior e inferior a la caja representan los valores $q_{75}+1.5(q_{75}-q_{25})$ y $q_{25}-1.5(q_{75}-q_{25})$ respectivamente. Los valores mayores o menores que estos límites se consideran valores extremos y se representan meddiante puntos.

Normalmente no se utilizan para una sola variable (es preferible el histograma) pero son muy utilizados para comparar las distribuciones de una variable cuantitativa para diferentes muestras caracterizadas por, por ejemplo, diferentes niveles de un factor (variable cualitativa) tal como se ve en la figura 3.11.

3.3.2 Correlación paramétrica

Se mide con la covarianza y el coeficiente de correlación:

$$COV_{xy} = \frac{\sum_{i=1}^{N}(x_i - m_x)(y_i - m_y)}{N} \tag{3.8}$$

$$r_{xy} = \frac{COV_{xy}}{s_x s_y} \tag{3.9}$$

Una de las ventajas del coeficiente de correlación es que es sencillo de interpretar ya que $1- <= r_{xy} <= 1$. Si se proxima a -1 existe una relación negativa entre las variables (cuando una aumenta la otra disminuye), si se aproxima a 1 existe una relación positiva (cuando una aumenta la otra también lo hace) y si se aproxima a cero no existe relación.

El coeficiente de correlación es uno de los estadísticos más utilizados por los geógrafos, pero su habitual mal uso genera muchos problemas que contribuyeron en su día a desprestigiar los métodos cuantitativos entre los propios geógrafos. Entre los errores habituales:

- Confundir correlación con causalidad. Si X está correlacionada con Y no se debe concluir necesariamente que X es causa de Y, puede ser que:
  - X influya sobre Y
  - Y influya sobre X
  - Exista una variable Z que influya sobre X e Y

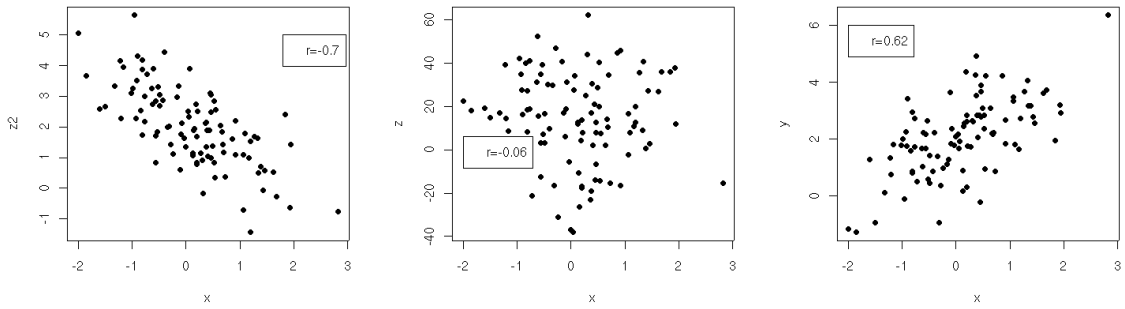


Figura 3.8: Distribución bivariante y coeficientes de correlación

- El coeficiente de correlación de Pearson debe utilizarse con datos procedentes de una distribución normal (histograma simétrico y de forma acampanada), emplearlo con variables que no cumplan esta condición puede dar lugar a conclusiones erroneas. En concreto la presencia de valores anormalmente extremos puede producir resultados contradictorios con la realidad. La figura 3.9 muestra dos ejemplos. Los valores extremos pueden resultar de mediciones erroneas o individuos no pertenecientes a la población que se pretendía estudiar.

3.3.3 Correlación no paramétrica

En el caso de que no podamos asumir que nuestros datos proceden de una distribución normal y sea imposible transformarlos. Puede utilizarse el **coeficiente de correlación de Spearman** que no se basa en las diferencias respecto a la media de los datos sino en la ordenación de los mismos. El siguiente ejemplo explicará el procedimiento.

En una serie de campos abandonados se ha medido la diversidad florística (f), se pretende determinar si el tiempo (t) desde el abandono influye o no sobre la diversidad. Los datos aparecen en la tabla3.1. En dicha tabla t_o es el número de orden del correspondiente valor de t, f_o es el número de orden del correspondiente valor de f y d es la diferencia entre los respectivos valores de orden. Determinar, con un nivel de significación de 0.05 si existe relación entre ambas variables.

La ecuación para calcular el coeficiente es:

$$r_s = 1 - \frac{6 \sum d^2}{n^3 - n} \tag{3.10}$$

en el ejemplo anterior:

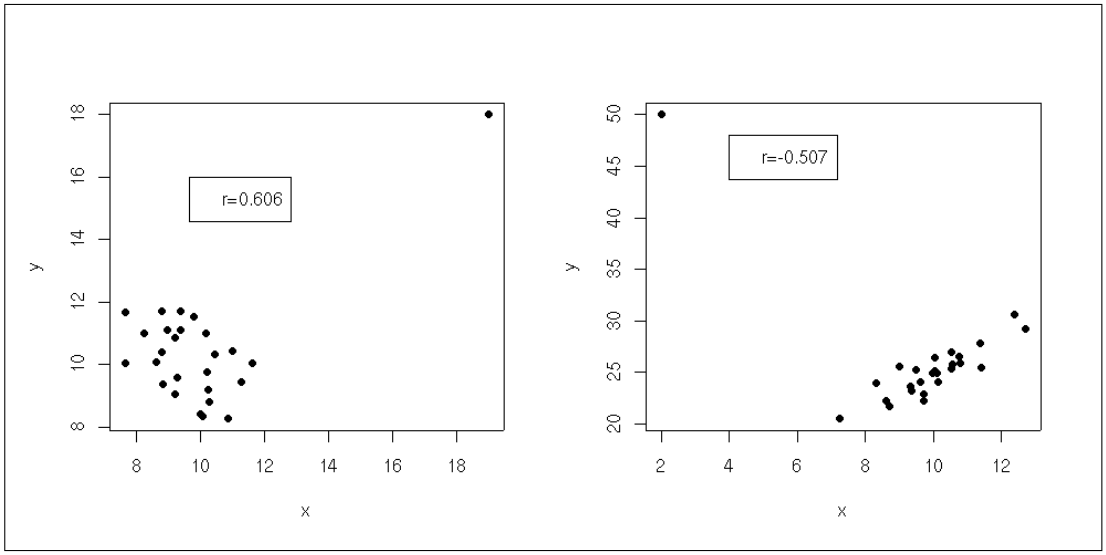


Figura 3.9: Ejemplo de coeficientes de correlación erroneos debido a lo inadecuado de las muestras

$$r_s = 1 - \frac{6x5.5}{343 - 7} = 0.902 \tag{3.11}$$

Los valores de r_s oscilan entre -1 y 1 igual que con el coeficiente de correlación de Pearson.

3.3.4 Con R

R dispone también de funciones para el análisis bi o multivariante, entre las más básicas:

- `cov(x,y)` para calcular la covarianza entre x e y
- `cor(x,y)` para calcular el coeficiente de correlación entre x e y

En R puedes calcular r_s con la función `cor` utilizando la opción `method="spearman"`. Esta opción puede utilizarse igualmente con la función `cor.test`.

Así la orden `cor(c(1,2,3,4,8,10,12),c(2,3,5,4,7,6,7),method="spearman")` devolverá el siguiente resultado:

```
        Spearman's rank correlation rho

data:  c(1, 2, 3, 4, 8, 10, 12) and c(2, 3, 5, 4, 7, 6, 7)
```

t	f	t_o	f_o	d	d^2
1	2	1	1	0	0
2	3	2	2	0	0
3	5	3	4	-1	1
4	4	4	3	1	1
8	7	5	6.5	-1.5	2.25
10	6	6	5	1	1
12	7	7	6.5	0.5	0.25

Table 3.1: Ejemplo del cálculo del coeficiente de correlación de Spearman

```
S = 5.5475, p-value = 0.005621
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.9009375
```

que nos indica que el valor de r_s es de 0.901 y que la probabilidad de obtener este valor, mayor, con una muestra de de ese tamaño, asumiendo que es cierta H_0 (es decir que no existe relación entre la diversidad florística y el tiempo desde el abandono de la parcela) es de un 0.0056, considerablemente inferior al nivel de significación pedido. Por tanto podemos aceptar la hipótesis alternativa H_1 de que si existe esa relación.

La representación gráfica de la relación entre dos variables puede hacerse con `plot` (figura 3.10) cuando x e y sean cuantitativas y con `boxplot` (figura 3.11) cuando x sea cualitativa e y cuantitativa.

El código de R para generar ambas figuras es:

```
plot(z,ss,cex=0.3,pch=16,xlab="Altitud (m)",ylab="Pendiente (%)",
main="Relación entre altitud y pendiente")
```

y

```
boxplot(z factor(f),names=c("Bosque","Cultivo","Matorral"),
xlab="Usos del suelo",ylab="Elevación en metros",
main="Distribución de altitudes por usos del suelo")
```

donde f es un vector numérico con los usos del suelo correspondientes a cada valor de z codificados con números del 1 al 3.

3.3.5 Ejercicios

1. Explicar por que son inadecuados los coeficientes de correlación obtenidos en la figura 3.9

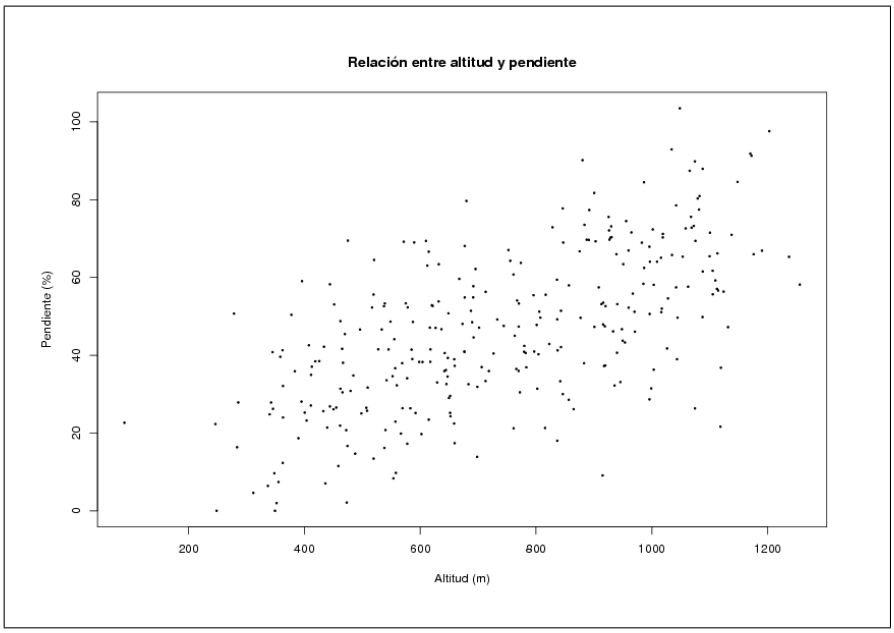


Figura 3.10: Gráfico plot

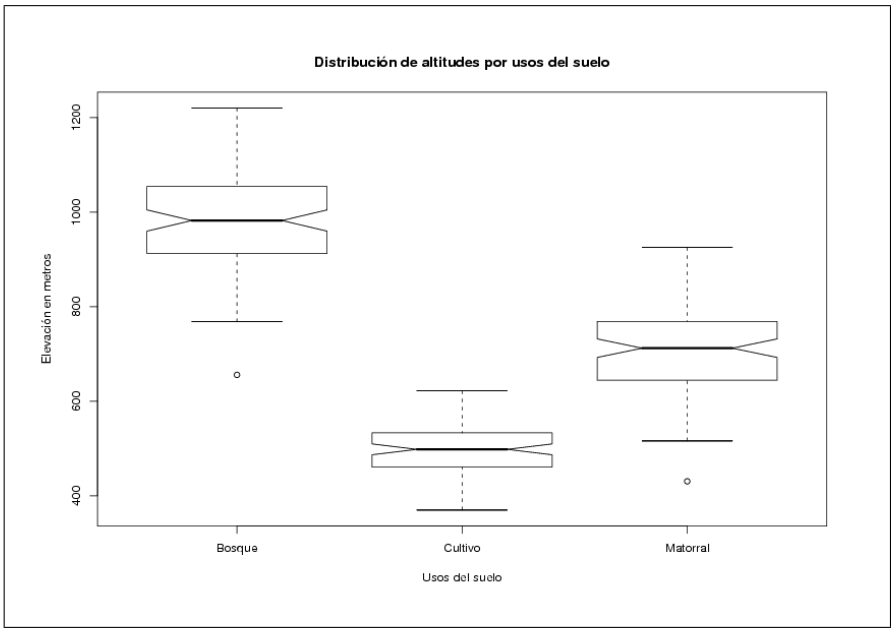


Figura 3.11: Gráfico boxplot

2. Obtener los estadísticos descriptivos básicos de las variables en el fichero base_datos.txt. ¿En que observatorio es mayor la precipitación?, ¿en cual más variable?

3. Representar algunas de las variables en el fichero base_datos.txt y relacionar el gráfico con el coeficiente de correlación calculado para las variables.

4. Calcula los coeficientes de correlación de Pearson y Spearman de algunas de las variables y trata de interpretar las diferencias.

Tema 4

Estadística inferencial. Caracterización de poblaciones a partir de muestras

Un jugador de dados del *far west* comprueba que su rival a obtenido dos seises y un cinco al comienzo de la partida. ¿Debe sacar su arma puesto que le están haciendo trampa o no?

El sentido común lleva a pensar que con sólo tres tiradas aún no podemos decidir si el dado está trucado o no. Para expresarlo en términos estadísticos, podríamos decir que el tamaño muestral es demasiado reducido para decidir si es cierto o no que:

$$Prob(n) = \frac{1}{6} \tag{4.1}$$

para valores de n entre 1 y 6, con un grado de significación estadística suficientemente elevado.

Por lo tanto, nuestro *tahur* decide, con buen criterio, esperar a verlas venir, o sea a tener un tamaño muestral más elevado.

En este caso, la decisión se basa en el conocimiento de cual debe ser la estructura de probabilidad del fenómeno analizado, es obvio que un dado sin trampa debe satisfacer la ecuación 4.1, a pesar de que los resultados pueden desviarse, en ocasiones de forma notable, del resultado esperado.

Antes de la aparición del cálculo electrónico, el material de prácticas de una clase de estadística podía consistir en varios dados con los que los alumnos debían elaborar largas tablas que permitieran comprobar o rechazar hipótesis. Hoy en día podemos simplificar el problema gracias a los programas de análisis de datos como R. Vamos a trabajar con una función sencillita, escribe:

```
> source(''funciones.R'')
>> dado()
```

La función `dado()` emula el lanzamiento de un dado, si no se le pasan parámetros se comporta como un dado normal de 6 caras que cumple la ecuación 4.1 pero podemos pasarle parámetros, por ejemplo:

```
> dado(n=100)
```

genera 100 tiradas de dado y

```
> tirada=dado(n=100,c=10)
```

genera 100 tiradas de un hipotético dado de 10 caras y las almacena en el objeto `tirada`. A continuación podemos tabular los resultados anteriormente obtenidos con la función `table`

```
> table(tirada)
```

Puedes comprobar que al aumentar el número de tiradas los valores de frecuencia recogidos se van acercando a $1/6 = 0.1666667$ (en el caso de un dado de 6 caras claro) pero es bastante improbable obtener exactamente este valor.

La conclusión a extraer es que los datos recogidos de un sistema sujeto a aleatoriedad cuyo comportamiento estadístico es perfectamente conocido pueden desviarse de forma notable de los esperados.

Ahora vamos a abordar este problema desde otro punto de vista. Un geógrafo obtiene del INM la serie de precipitación anual de un pluviómetros situado en la zona de estudio de su tesina. Desgraciadamente el pluviómetro tiene poco tiempo y la serie es corta:

```
85.6 445.2 360.3 195 311 141 636.2 171 466.9 291.9
```

Analizando estos datos se obtiene que $m_p = 310.4$, $s_p = 171.1$, el histograma es el que aparece en la figura 4.1, y por ejemplo la probabilidad de no llegar un año a 100 mm es de 1/10=0.1 o sea el 10 %.

¿Podríamos asumir que la estructura de probabilidad del fenómeno (media y desviación típica de la precipitación anual) es igual que la de la muestra obtenida? Evidentemente no, sería lo mismo que si el jugador del primer ejemplo se hubiese liado a tiros tras las tres primeras tiradas. La diferencia es que en aquel caso conociamos cual debía ser la estructura de probabilidad pero no es así en este caso.

Sin embargo en muchas ocasiones se tiende a exagerar la importancia de los datos obtenidos aunque los tamaños muestrales sean reducidos. Esto es especialmente frecuente entre los geógrafos que tienden hacia una ingenua sobrevaloración del empirismo en parte por desconocimiento de las técnicas de análisis de datos cuya utilización permite una mejor caracterización de la realidad que los datos por si mismos.

Ejercicios

- Realiza el experimento con la función `dado()` con tiradas de 100, 500, 1000, 5000, 10000, 50000, 100000 y 500000 dados.
- Inventa y calcula un índice que muestre para cada tirada lo cerca o lo lejos que esta de la hipótesis reflejada en la ecuación 4.1.
- Haz un gráfico que muestre como evoluciona el índice con el aumento del número de dados por tirada. ¿Qué conclusiones puedes sacar?

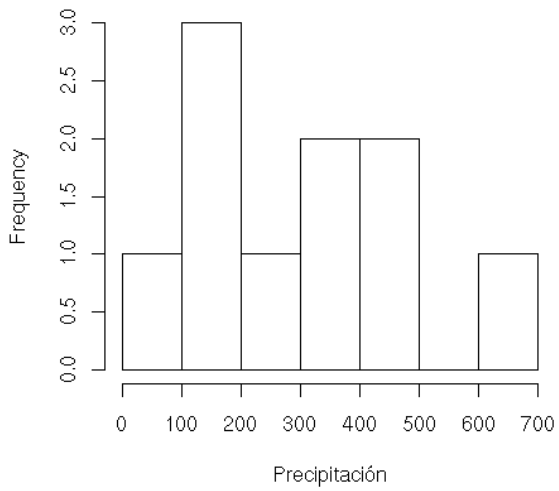


Figura 4.1: Histograma de precipitación

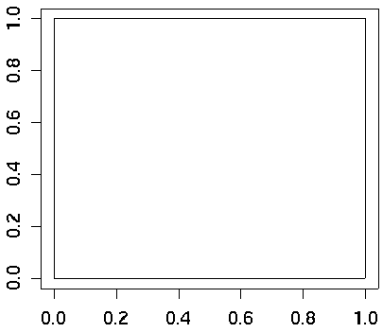
4.1 Probabilidad

La estadística tiene su base en la teoría de probabilidades que asigna a los resultados de un experimento un valor de probabilidad de ocurrencia. La palabra *experimento* se utiliza aquí en sentido muy amplio, que incluye cosas como lanzar un dado o preguntarle a alguien por la calle su edad.

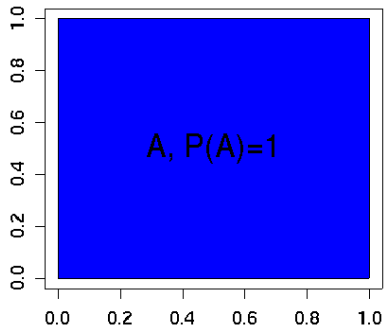
Para mejor entender la probabilidad y sus leyes, vamos a utilizar un procedimiento gráfico (figura 4.2) en el que la probabilidad de un suceso se representará mediante un cuadrado en un espacio de área 1 (4.2:a), de manera que la probabilidad del suceso será igual al área de un cuadrado. Así un suceso seguro ocupará todo el espacio (4.2:b), mientras que un suceso representedo por un cuadradado de 0.4 de lado tendrá una probabilidad de 0.16 (figura 4.2:c).

La probabilidad cumple una serie de propiedades:

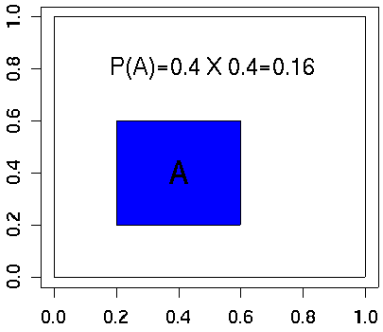
- La probabilidad está siempre entre 0 y 1
  $0 \leq P \leq 1$
- La probabilidad del suceso imposible es 0
  $P(0) = 0$
- La probabilidad del suceso seguro es 1
  $P(E) = 1$



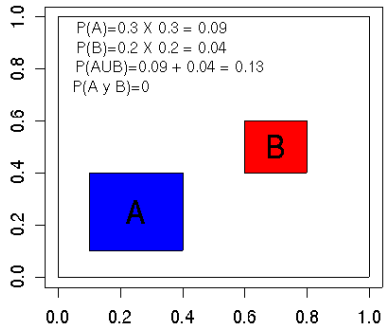
a)



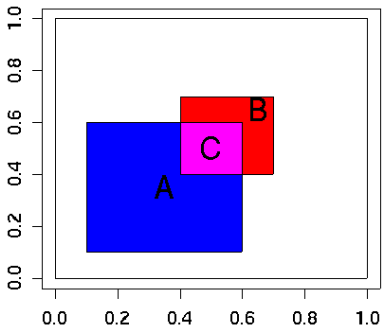
b)



c)



d)



e)

Figura 4.2: Representación gráfica de la probabilidad y sus leyes

- La probabilidad de que se produzca o no se produzca el suceso A es 1

  $P(A + \bar{A}) = P(A) + P(\bar{A}) = 1$

- La probabilidad de que se produzca un suceso A o un suceso B es igual a la suma de sus probabilidades menos la probabilidad de que se produzcan al mismo tiempo:

  $P(A \bigcup B) = P(A) + P(B) - P(A \bigcap B)$

  La figura 4.2:e muestra como si sumamos $P(A)+P(B)$ habremos contado dos veces $P(C) = P(A \bigcap B)$ por lo que es necesario restar una vez dicha cantidad

  Por tanto si dos sucesos A y B son incompatibles (no se pueden producir a la vez) $P(A \bigcup B) = P(A) + P(B)$ (figura 4.2:d)

- La probabilidad de que se produzca un suceso A y un suceso B es igual a la probabilidad de que se produzca A sabiendo que se ha producido B $P(A|B)$ por la probabilidad de B, y también es igual a la probabilidad de que se produzca B sabiendo que se ha producido A $P(B|A)$ por la probabilidad de A.

$$P(A \bigcap B) = P(A|B)P(B) = P(B|A)P(A) \tag{4.2}$$

  La figura 4.2:e muestra este concepto, partiendo de la concepción geométrica anteriormente expuesta:

  - $P(A \bigcap B) = 0.2x0.2 = 0.04$
  - $P(A|B)P(B) = [(0.2x0.2)/(0.3x0.3)]x0.3x0.3 = 0.04$
  - $P(B|A)P(A) = [(0.2x0.2)/(0.5x0.5)]x0.5x0.5 = 0.04$

Ejercicios

1. ¿Cual será la probabilidad de sacar de una baraja española (40 cartas) una copa?

2. ¿Cual será la probabilidad de sacar una copa o una espada? ¿y de sacar una copa y una espada? Justifícalo en términos de probabilidad.

3. ¿y una carta menor que el 5 (as=1)?

4. ¿y una copa menor que 5?

5. ¿Cual sería la probabilidad de obtener primero una sota y luego un rey?

4.2 Inferencia a partir de los estadísticos descriptivos

Podemos calcular con exactitud los estadísticos de una muestra, pero no los parámetros poblacionales equivalentes, en todo caso podemos dar una aproximación. Esta aproximación se basa en un intervalo de valores centrado en el valor del estadístico. La anchura de este intervalo dependerá de la probabilidad que que queramos tener de no equivocarnos y del tamaño de la muestra utilizada para calcular el estadístico. En todos los casos se parte del **error estándar (o error típico)** del estadístico que se suelo escribir SE.

Antes de analizar los errores estándar es conveniente entender un poco mejor la distribución normal (figura 5.2). Podemos considerar que una variable sigue una distribución normal si su histograma cuando el tamaño muestral tiende a infinito, y por tanto las barras del histograma pueden hacerse tan pequeñas como queramos, tiende a la figura 5.2:e a medida que el número de barras (NB) aumenta.

En esta figura puede verse además como las barras del histograma representan, no el número total de casos sino su frecuencia o probabilidad, es decir el número de casos representados por la barra partido por los casos totales, de manera que la suma de todas las barras de cualquiera de los histogramas la figura 5.2 es igual a uno. Siguiendo el mismo razonamiento, el área bajo la curva normal en 5.2:e también es 1. Pero además podemos obtener el área de la curva entre dos valores cualesquiera:

- El área de la curva normal entre -1 y 1 es 0.682 (área verde en la figura 5.2:f).
- El área de la curva normal entre -2 y 2 es 0.954 (áreas verde y azul en la figura 5.2:f).
- El área de la curva normal entre -3 y 3 es 0.997 (áreas verde, azul y roja en la figura 5.2:f).

Estos valores concretos van a ser importantes para estimar los valores, entorno al valor de un estadístico entre los que, con determinada probabilidad, podemos considerar que está el valor del parámetro correspondiente.

La figura 5.2 muestra el caso de una distribución normal con media 0 y desviación típica 1, lo que se denomina una **normal tipificada**, pero lo visto anteriormente puede aplicarse también a distribuciones normales con cualquier valor de media y de desviación típica.

4.2.1 Con R

Con R podemos obtener la probabilidad de obtener cualquier valor x procedente de una población que sigue una distribución normal:

```
 > dnorm(3,mean=0,sd=1)
[1] 0.004431848
```

también podemos obtener la probabilidad de obtener un valor igual o inferior a x:

```
 > pnorm(3,mean=0,sd=1)
[1] 0.9986501
```

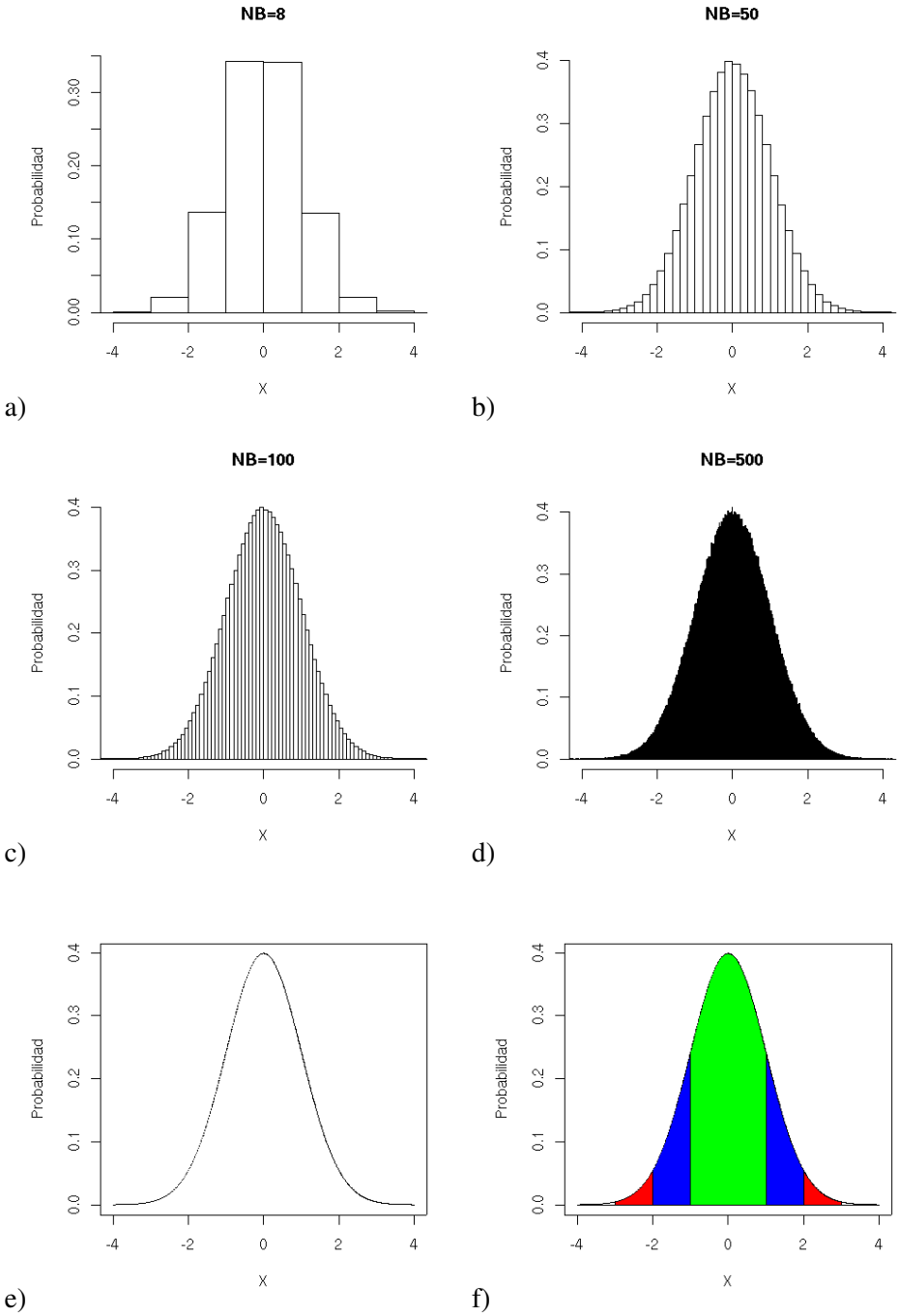


Figura 4.3: Distribución normal

A partir de esta función podemos obtener la probabilidad de que x esté entre dos valores determinados `A` y `B`, ya que:

$$P(A < X \leq B) = P(X \leq B) - P(X \leq A) \tag{4.3}$$

Así si queremos calcular cual es el área bajo la curva normal tipificada entre -1 y 1 procederíamos así:

```
> pnorm(1,mean=0,sd=1) - pnorm(-1,mean=0,sd=1)
[1] 0.682
```

como ves el resultado coindice con el expuesto anteriormente.

Por otra parte podemos obtener el valor K cuya probabilidad de que $x \leq K$ sea igual a un valor dado:

```
> qnorm(0.9986501,mean=0,sd=1)
[1] 3
```

Observa que esta función es la inversa de `pnorm` y nos va a permitir deducir entre que par de valores (centrados en la media) se encuentra un determinado porcentaje del área bajo la curva normal:

Si queremos un área X debemos tener en cuenta que vamos a dejar un área de $(1 - X)/2$ a la derecha de la curva y un área de $(1-X)/2$ a la izquierda. Por tanto debemos calcular cuales son los valores cuya probabilidad de no ser superados son $(1 - X)/2$ y $1 - ((1 - X)/2)$.

Suponemos que X=0.95, por tanto las probabilidades a la derecha e izquierda serían 0.025 y 0.975 respectivamente y los correspondientes valores se obtendrían:

```
> qnorm(0.025,mean=0,sd=1)
[1] -1.959964
> qnorm(0.975,mean=0,sd=1)
[1] 1.959964
```

En realidad, dada la simetría de la curva normal, basta con calcular un valor y cambiar el signo para obtener el otro (ver la figura 4.4).

Ejercicios

- Calcula la probabilidad de obtener un valor entre 2 y 3.5 de una población que sigue una distribución normal con media 2 y desviación típica 1.
- Calcula que valor tiene una probabilidad de no ser superado del 99% en una distribución normal con media 2 y desviación típica 1. ¿Cual sería la probabilidad de ser superado?
- Calcula en que intervalo de valores, centrado en la media, se situará el 99% de los valores de una variable que tiene media 0 y desviación típica 1

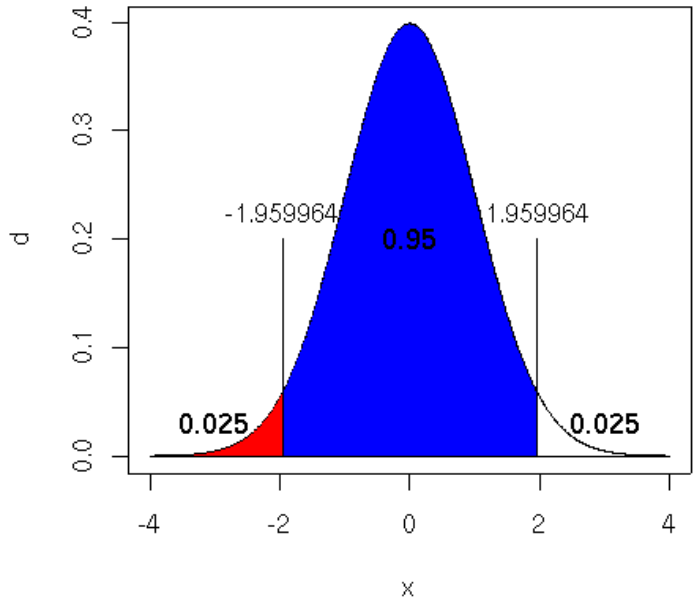


Figura 4.4: Cálculo de los límites de un área bajo la normal dada

4.2.2 Error estándar de la media

Si asumimos que una muestra procede de una población con distribución normal, podemos confiar en que la media muestral se aproximará a la media poblacional cuanto más grande sea el tamaño muestral. Pero el grado de aproximación dependerá también de la varianza de la población (aproximada por la varianza de la muestra).

De esta manera podemos asumir que el error de la media muestral disminuye con el tamaño muestral y aumenta con la varianza de la muestra. Por tanto la ecuación para calcular el error estándar o error típico de la media es:

$SE_m = \sqrt{\frac{s^2}{n}}$

De esta manera no sabremos cual es realmente el valor de la media poblacional (μ) pero podemos estar razonablemente seguros de que:

$m - SE_m \leq \mu \leq m + SE_m$

Pero, puesto que la inferencia estadística es siempre una cuestión de probabilidad, podemos establecer que:

1. $\mu = m \pm SE_m$, con una probabilidad de 0.682, la probabilidad de equivocarnos es 0.318
2. $\mu = m \pm 2SE_m$, con una probabilidad de 0.954, la probabilidad de equivocarnos es 0.046
3. $\mu = sm \pm 3SE_m$, con una probabilidad de 0.997, la probabilidad de equivocarnos es 0.003

Los anteriores enunciados podrían reformularse así:

1. Hay una probablidad de 0.682 de que μ esté entre $m - SE_m$ y $m + SE$, y un 0.318 de que no
2. Hay una probablidad de 0.954 de que μ esté entre $m - 2SE_m$ y $m + 2SE$, y un 0.046 de que no
3. Hay una probablidad de 0.997 de que μ esté entre $m - 3SE_m$ y $m + 3SE$, y un 0.003 de que no

Estos intervalos se denominan **intervalos de confianza** para una determinada probabilidad, que nos da una medida del grado de incertidumbre de la media calculada. Lógicamente cuanto mayor queramos hacer la probabilidad de no equivocarnos mayor será el intervalo de confianza y más inutil el resultado.

Habrás observado que los valores que se han dado son los mismos que utilizamos para calcular áreas bajo la curva normal en función de la desviación típica.

Así podemos establecer un intervalo de confianza que puede calcularse como:

$\mu = m \pm Z_{\alpha/2}\sqrt{\frac{s^2}{n}}$

donde $Z_{\alpha/2}$ es el área bajo la normal tipificada (es decir `qnorm(a/2,mean=0,sd=1)` en R) y α el **nivel de significación** o grado de incertidumbre que estamos dispuestos a asumir, es decir la probabilidad de equivocarnos o 1 menos la probabilidad de acertar.

Un ejemplo

En el archivo `murcia_anual.txt` que puedes bajar de la página web de la asignatura hay una simulación de datos de precipitación anual en Murcia (no son datos reales). A partir de estos datos determina cual es el la precipitación media poblacional con un nivel de significación de 0.05.

1. En primer lugar cargamos los datos: `murcia=read.table("murcia_anual.txt")`
2. Calculamos los estadísticos muestrales y el error típico: `m=mean(murcia);s=sd(murcia);se=s/sqrt(le`
3. Obtenemos el valor de $Z_{\alpha/2}$ como el valor absoluto de `Z=qnorm(a/2,mean=0,sd=1)` (donde `a`=α)
4. Finalmente establecemos los intervalos de confianza:

   ```
   mi=m-z*se;ms=m+z*se
   ```

   donde `mi`=293.3 es el valor inferior del intervalo de confianza y `ms`=308.7 el valor superior, luego: $293.3 \leq \mu \leq 308.7$ con un nivel de significación de 0.05, es decir que la probabilidad de equivocarnos y que μ sea inferior o superior a dichos valores es sólo del 5%.

4.2.3 El error estándar de la desviación típica

La ecuación para calcular el error estándar de la desviación típica es:
$SE_s = \frac{s}{\sqrt{2n}}$
de manera que:

1. $\sigma = s \pm SE_s$, con una probabilidad de 0.682
2. $\sigma = s \pm 2SE_s$, con una probabilidad de 0.954
3. $\sigma = s \pm 3SE_s$, con una probabilidad de 0.997

Puede obtenerse el intervalo de confianza para un grado de incertidumbre dado de la misma forma que en el caso de la media.

4.2.4 El error estándar del coeficiente de sesgo

La ecuación para calcular el error estándar del coeficiente de sesgo es:
$SE_g = \sqrt{\frac{6}{n}}$
de manera que:

1. $\gamma = g \pm SE_g$, con una probabilidad de 0.682
2. $\gamma = g \pm 2SE_g$, con una probabilidad de 0.954
3. $\gamma = g \pm 3SE_g$, con una probabilidad de 0.997

Si el intervalo de confianza del coeficiente de sesgo incluye al 0 (por ejemplo está entre -0.5 y 5) podemos asumir que la distribución es normal. La comprobación de la hipótesis $\gamma_x = 0$ es importante de cara a saber si la muestra sigue o no una distribución normal ya que si no la sigue no se podrán utilizar correctamente ni los errores estándar ni muchas de las técnicas que se revisarán posteriormente.

Normalmente se sugire que si no podemos aceptar que $\gamma_x = 0$ debe trabajarse con una transformación de x:

- Si g es ligeramente positivo: $y = \sqrt{x}$
- Si g es ligeramente negativo: $y = x^2$
- Si g es bastante mayor que 0: $y = log(x + C)$
- Si g es bastante menor que 0: $y = log(K - x)$
- Si g es mucho mayor que 0: $y = 1/(x + C)$
- Si g es mucho menor que 0: $y = 1/(K - x)$

Lo mejor es dibujar el histograma de las variables transformadas, y volver a calcular el coeficiente de sesgo con las variables transformadas, para determinar cual es la más se asemeja a una distribución normal. Si conseguimos que la variable transformada siga una distribución normal podremos aplicar las técnicas paramétricas que sólo deben aplicarse a distribuciones normales.

Un ejemplo

Del fichero `base_datos.txt` vamos a extraer la precipitación de octubre en Zarzadilla de Totana (precZ) y a calcular su coeficiente de sesgo con un nivel de significación de 0.01.

1. `datos=read.table("base_datos.txt",header=T);P=datos$precZ`
2. `source("funciones.R");g=sesgo(p);se=sqrt(6/length(p))`
3. `z=qnorm(0.005,mean=0,sd=1)`
4. `gi=g-z*se;gs=g+z*se`

donde `gi`=0.4 es el valor inferior del intervalo de confianza y `ms`=2.39 el valor superior, luego: $0.4 \leq \mu \leq 2.39$ con un nivel de significación de 0.01, es decir que la probabilidad de equivocarnos y que γ sea inferior o superior a dichos valores es sólo del 1%, lo que implica que debemos rechazar la hipótesis de que $\gamma = 0$.

En realidad esto lo podríamos haber comprobado representando el histograma de `p` (figura **??**). Deberíamos por tanto escoger una transformación de variable.

4.2.5 Errores estándar de un conjunto de proporciones

La ecuación para calcular el error estándar de una proporción es:

$$SE_p = \sqrt{\frac{pq}{n}} \tag{4.4}$$

donde p es la proporción (en tantos por uno), $q = 1 - p$ y n es el tamaño muestral.

así que al igual que antes:

1. $\pi_{pob} = p_m \pm SE_p$, con una probabilidad de 0.682
2. $\pi_{pob} = p_m \pm 2SE_p$, con una probabilidad de 0.954
3. $\pi_{pob} = p_m \pm 3SE_p$, con una probabilidad de 0.997

y de nuevo:

$\pi_{pob} = p_m \pm Z_{\alpha/2} SE_p$

4.2.6 Aplicación de los errores típicos

Las ecuaciones de los diversos errores estándar pueden utilizarse para determinar el tamaño adecuado de la muestra para calcular un estadístico con un determinado error típico a partir de una primer muestreo tentativo para obtener la desviación típica.

Por ejemplo en el caso de necesitar una estimación de la media con un error típico de 1, si se han tomado 10 puntos y el valor de la desviación típica es $s = 10$, resulta que el error típico de la media es:

$$SE_m = \frac{s}{\sqrt{n}} = \frac{10}{4.47} = 2.23 \tag{4.5}$$

en este caso el error de la media obtenido es mayor que el deseado, pero podemos evaluar el tamaño muestral necesario despejando n de la ecuación del error típico y haciendo $SE_m = 1$:

$$n = (\frac{s}{SE_m})^2 = (\frac{10}{1})^2 = 100 \tag{4.6}$$

Si al recalcular el error típico con una muestra de tamaño 100 resulta que $SE_m = 1$ será por que s habrá aumentado, por lo que habrá que utilizar el nuevo valor de desviación típica para hacer una nueva estimación de n.

Ejercicios

- Calcula la probabilidad de obtener un valor entre 2 y 3.5 de una población que sigue una distribución normal con media 2 y desviación típica 1.
- Calcula la media y desviación típica de la precipitación en librilla con un nivel de significación de 0.01.
- Calcula el coeficiente de sesgo de la precipitación en Librilla. ¿Puedes asumir con un nivel de significación de 0.05, que procede de una distribución normal? Si no es así calcula una variable transformada que cumpla esta hipótesis. Compruébalo pintando el histograma y volviendo a calcular el coeficiente de sesgo de la variable transformada.

4.3 Inferencia acerca de la correlación

Anterioremente se ha visto como el coeficiente de correlación adopta valores entre -1 y 1 y cuanto más se aproxime a estos más clara es la relación. El problema que se plantea es que hacer cuando el valor de r esté en torno a -0.5 o 0.5.

¿Cómo decidir si en la población estudiada, de la que hemos extraido una muestra para calcular r, existe o no una relación entre las variables?

En estos casos debe calcularse la probabilidad de obtener ese resultado por casualidad asumiendo que no hay relación entre las variables. Si esta probabilidad es muy pequeña asumiremos que r es significativamente distinto de cero y por tanto si existe esa relación.

$$z = |r|\sqrt{(n)} \tag{4.7}$$

A partir de este índice y con R podemos calcular:

```
p=1-pnorm(z)
```

que nos dará la probabilidad de que siendo el coeficiente de correlación cero, hubiesemos obtenido por azar el valor `r`. Un valor suficientemente pequeño nos llevará a pensar que el coeficiente de correlación será distinto de cero. Generalmente se decide cual será el valor umbral antes de llevar a cabo el análisis.

Ejercicios

- Se dispone de dos muestras de precipitación anual, la muestra A procede de un observatorio en el interior, mientras que la muestra B procede de un observatorio en la costa.

  A=c(305,189,244,198,254,267,253,335,335,464,236,321, 49,387,224,205,346,331,112,405,277,331,95,90,476)

  B=c(193,524,668,136,225,572,289,627,609,159,673,638,169,724,293,404,520,494,85,685,394,552,231,360,396)

  A pesar de que ambas muestras tienen propiedades estadísticas diferentes se plantea la hipótesis de que muestran cierta correlación. Determinar si esto es así con un nivel de significación de 0.01 y de 0.05.

4.4 Contraste de hipótesis

En el apartado anterior hemos visto como inferir el valor de ρ (coeficiente de correlación poblacional) a partir de r y al mismo tiempo estabamos verificando la hipótesis de que $\rho = 0$ o $\rho \neq 0$.

Cuando observamos un fenómeno, podemos establecer diversas hipótesis acerca de su naturaleza. Para poder decidir si nuestra hipótesis es cierta o no (con cierto grado de seguridad) deberemos:

- Determinar cuales son las variables (**variables explicativas** y **variables respuesta**) que permiten **modelizar** el fenómeno con un nivel de simplicidad adecuado.
- Establecier un procedimiento de **muestreo**.
- Medir las variables seleccionadas en los **individuos** que componen la **muestra**.
- Hacer un **análisis exploratorio**, generalmente gráfico, para hacernos una primera idea del comportamiento de las variables.
- Establecer cual es la **Hipótesis nula (H_0)** y la **hipótesis alternativa (H_a)** y el **nivel de significación** que daremos por bueno para decidir cual se cumple.
- Aplicar un **análisis estadístico** adecuado. La elección de un tipo u otro depende de la naturaleza de los datos implicados y del tipo de hipótesis que queremos verificar.
- Decidir, en función de los resultados del análisis estadístico y del nivel de significación establecido *a priori*, si se cumple la hipótesis nula o la alternativa.

En algunos casos, si aceptamos la hipótesis alternativa podremos generar un **modelo empírico o estadístico** que me permita hacer estimaciones de los valores de las variables respuesta en un individuo conociendo los valores de las variable explicativas.

Un modelo estadístico también puede ser un conjunto de parámetros que resuma las características de un fenómeno y me permita **simular** valores de las variables que describen dicho fenómeno:

- Sabiendo como se comporta un dado puedo simular tiradas de dado
- Sabiendo cuales son las características de la precipitación en un observatorio podemos generar series de precipitación.

Antes se ha afirmado que todo análisis estadístico comienza con un fenómeno que nos llama la atención. Este fenómeno suele ser del tipo:

- Dos muestras parecen tener características diferentes:
  - La precipitación media es diferente entre dos observatorios cercanos

  - El dado con el que quieres jugar está trucado (sus características son diferentes a las de un dado convencional) En estos casos la hipótesis nula sería que no hay diferencias, la hipótesis alternativa que si las hay.

- Una muestra parece ajustarse especialmente bien a una función de distribución (tema 5)
- Un conjunto de variables parecen tener relación entre ellas (tema **??**)
  - La precipitación se incrementa con la altitud
  - La temperatura desciende con la altitud
  - El contenido en materia orgánica de un suelo depende del uso de ese suelo

La hipótesis alternativa es la que queremos comprobar, la hipótesis nula sería la contraría, es decir:

- La precipitación NO se incrementa con la altitud
- La temperatura NO desciende con la altitud
- El contenido en materia orgánica de un suelo NO depende del uso de ese suelo

Puesto que para decidir cual es la hipótesis cierta vamos a analizar una muestra de tamaño necesariamente reducido **nunca podremos estar seguros al 100% de cual es la hipótesis cierta**, lo que si podemos decidir es con que porcentaje queremos estar seguros o, lo que es lo mismo, cual es el **nivel de significación** que queremos para tomar nuestra decisión.

4.4.1 Contraste de hipótesis respecto a la media

Si comparamos datos de precipitación anual en dos localidades, podemos plantear las siguientes hipótesis:

- H_0: Ambas localidades tienen la misma precipitación media
- H_a: Las precipitaciones medias son diferentes.

Debemos distinguir entre **media poblacional** (la que realmente tiene el fenómeno y que no podemos calcular sino como mucho estimar) y la **media muestral** (la que podemos calcular a partir de la muestra).

Aunque las medias poblacionales fueran iguales, es posible que las medias muestrales sean distintas debido a diversos factores al azar.

Para comprobarlo podemos utilizar el test **t de Student** que se basa en el estadístico:

$$t = \frac{|m_1 - m_2|}{\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}\frac{n_1+n_2}{n_1 n_2}}} \tag{4.8}$$

A cada valor de t se le asocia una probabilidad que es una estimación de la probabilidad de que un resultado en particular o un resultado aún más extremo pueda deberse al azar (en el caso de que la hipótesis nula sea cierta).

En el caso que nos ocupa sería la probabilidad de obtener un valor del estadístico *t de Student* igual o mayor al que hemos calculado en el caso de que la hipótesis nula sea cierta.

Por tanto todo estadístico debe venir acompañado de su valor *p*. El nivel de significación sería $100(1-p)$. Por ejemplo a un valor $p = 0.05$ le corresponde un nivel de significación del 95%.

4.4.2 Contraste de hipótesis respecto a l amedia con datos pareados

Si ambas muestran tienen el mismo tamaño y se pueden establecer parejas siguiendo algún criterio entre los datos de una y otra (por ejemplo datos de precipitación anual procedentes de los mismos años) es preferible utilizar el test **t de Student para datos pareados**.

Se trata de obtener una variable d con las diferencias entre las muestras, a partir de esta nueva variable, t se calcula como:

$$t = \frac{m_d}{\sqrt{\frac{\sum d^2 - \frac{(\sum d)^2}{n}}{n(n-1)}}} \tag{4.9}$$

donde n es el tamaño de cada una de las muestras.

4.4.3 Con R

El test *t de Student* se implementa con la función `t.test`:

```
> x=rnorm(100,mean=20)
> y=rnorm(50,mean=10)
> t.test(x,y)
```

La salida de este test será similar a esta:

```
        Welch Two Sample t-test

data:  x and y
t = 58.4103, df = 105.034, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  9.41693 10.07873
sample estimates:
mean of x mean of y
 20.02158  10.27375
```

en la que $t = 58.41$ con una p asociada de $2.2 * 10^{-16}$ es decir que la probabilidad de haber obtenido un valor de $t \geq 58.41$ por azar siendo las medias iguales es prácticamente cero con lo que se rechaza la hipótesis nula y se acepta la hipótesis alternativa de que ambas medias son diferentes ($m_x = 20.02$, $m_y = 10.27$).

Un uso alternativo de esta función sería comparar una muestra con un valor de media poblacional (y ya de paso veremos como cambiar el nivel de significación.

```
> x=rnorm(100,mean=10,s=1)
> t.test(x,mu=10,conf.level=0.99)
```

En este caso la respuesta será algo similar a:

```
        One Sample t-test

data:  x
t = -1.4769, df = 99, p-value = 0.1429
alternative hypothesis: true mean is not equal to 10
99 percent confidence interval:
  9.637692 10.101497
sample estimates:
mean of x
 9.869595
```

Donde lo interesante del resultado está de nuevo en los valores de t y p, que no permite rechazar la hipótesis nula, y en el intervalo de confianza para un nivel de significación del 99% ($9.64 \leq m_x \leq 10.1$).

Si lo que queremos es calcular el test t para datos pareados lo debemos indicar en la llamada a la función:

```
> x=rnorm(100,mean=20)
> y=rnorm(100,mean=20)
> t.test(x,y,paired=T)
```

El resultado será algo similar a:

```
        Paired t-test

data:  x and y
t = 0.6454, df = 99, p-value = 0.5202
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.1928571  0.3787938
sample estimates:
mean of the differences
             0.09296834
```

que nos da el valor de t, la media de las diferencias y el intervalo de confianza de la diferencia de las medias para un nivel de significación de 0.05

Existe un test similar para comparar varianzas (`var.test`):

```
> x=rnorm(100,mean=10,s=7)
> y=rnorm(50,mean=10,s=10)
> var.test(x,y)

        F test to compare two variances

data:  x and y
F = 0.443, num df = 99, denom df = 49, p-value = 0.000619
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.2662847 0.7077147
sample estimates:
ratio of variances
         0.4430457
```

Así como alternativas no paramétricas como el test U de Mann-Whitney.

4.5 Recapitulación

Habrás visto que las técnicas de inferencia estadística se basan en propiedades de la distribución normal, por tanto se asumen que los datos proceden de una distribución de ese tipo. Si no es así no se podrán utilizar salvo que hagamos una trasnformación de la variable.

Por otro lado, te habrás dado cuenta de que en muchas ocasiones el aceptar o no una determinada hipótesis depende del nivel de significación escogido, lo que lleva a dos consideraciones:

1. No hay que hacer trampas, el nivel de significación se establece *a priori*, no debe elegirse en función de los resultados
2. Es necesario tener alguna guía acerca de cual es el nivel de significación adecuado. En general depende de lo que nos estemos jugando. A continuación van 3 ejemplos:
   - Si al jugador de dados rechazar la hipótesis nula le supone tener que enfrentarse a tiros querrá estar muy seguro (nivel de significación muy bajo o lo que es lo mismo probabilidad de acertar muy alta), si sólo va a ser una bronca a gritos se conformará con un nivel de significación algo mayor.
   - Si la aceptación de la hipótesis alternativa conlleva iniciar un proyecto muy costoso querremos estar muy seguros, si va a ser algo corto y barato nos conformaremos con menos seguridad.
   - Al determinar si un río se va a desbordar o no, si en el llano de inundación hay sólo campos de cultivo el nivel de significación puede ser más bajo que si hay un hospital.

Tema 5

Inferencia acerca de la distribución

En el capítulo anterior se ha visto como inferir propiedades de una serie de parámetros poblacionales a partir de los estadísticos muestrales. Ahora vamos a estimar las características de la distribución de la que proceden los datos y cuya forma puede intuirse observando el histograma (siempre que el tamaño muestral sea suficientemente alto).

Un modelo de distribución se define por su función de densidad:

$$f(x) = prob(x) \tag{5.1}$$

y su función de distribución que es la función de densidad acumulada. En el caso de una distribución para variable discreta sería:

$$F(x) = prob(i \leq x) = \sum_{i=-\infty}^{x} f(i) \tag{5.2}$$

y en el caso de variable continua sería:

$$F(x) = prob(i \leq x) = \int_{i=-\infty}^{x} f(i)di \tag{5.3}$$

De la definición de función de densidad derivan los conceptos de **esperanza matemática** y **varianza** como generalizaciones de la media y varianza muestrales. La esperanza matematica para variables discretas se define cómo:

$$E[x] = \sum_{i=-\infty}^{\infty} xf(i)dx \tag{5.4}$$

mientras que para variables continuas es:

$$E[x] = \int_{i=-\infty}^{\infty} xf(i)dx \tag{5.5}$$

La varianza por su parte sería:

$$Var[x] = E[x^2] - (E[x])^2 \tag{5.6}$$

Existen diferentes modelos de distribución definidos por su función de densidad (que es una ecuación que depende de un conjunto de parámetros); la esperanza y varianza (que pueden calcularse de forma sencilla a partir de los parámetros de la función de densidad.

En el caso de variables aleatorias discretas, la función de distribución es un simple sumatorio, en el caso de variables continuas la función de distribución puede estar definida (se calcula integrando la función de densidad) o no. En este último caso puede calcularse con métodos numéricos con un ordenador.

A continuación se estudiarán algunos modelos de distribución para variables aleatorias discretas o continuas de utilidad en Geografía Físisca y en general en Ciencias Medioambientales. Ya has visto como se utiliza la distribución normal para hacer tests estadísticos, existen otras diseñadas con el mismo fin como la χ^2 (chi cuadrado) o la F de Snedecor que se verán más adelante.

5.1 Modelos de distribución para variables discretas

- Modelo uniforme

  Válido para el ejemplo del dado, todos los posibles valores tienen la misma probabilidad de ocurrencia, sus parámetros son el valor mínimo (min) y el valor máximo (max), su función de densidad es:

  $$f(k) = \frac{1}{1 + max - min} \tag{5.7}$$

  la esperanza y la varianza se definen como:

  $$E[k] = \frac{max + min}{2} \tag{5.8}$$

  $$Var[k] = (max - min)(max - min + 2)/12. \tag{5.9}$$

- Modelo binomial

  Se utiliza para determinar la probabilidad de ocurrencia de un suceso k veces de un total de n, sabiendo que la probabilidad de ocurrencia es p y la de no ocurrencia $q = 1 - p$. Por ejemplo puede utilizarse para evaluar la probabilidad de obtener k caras en un total de n lanzamientos de moneda.

  Se ha utilizado también para analizar la probabilidad de que se de un determinado número (k) de días de lluvia al cabo de un mes ($n = 30$) conociendo la probabilidad (p) de día lluvioso.

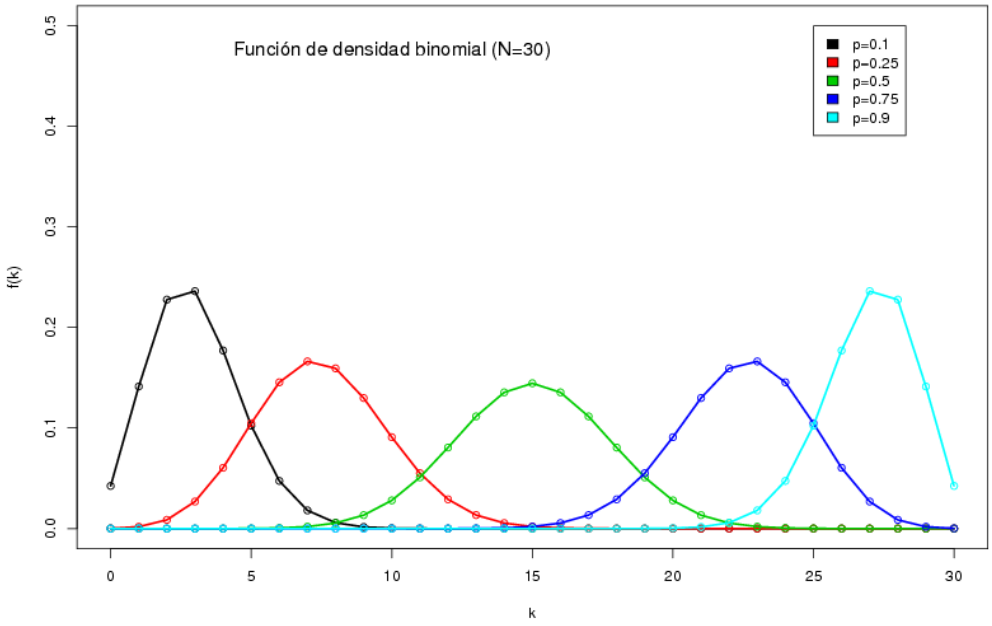


Figura 5.1: Ejemplos de Función de densidad binomial

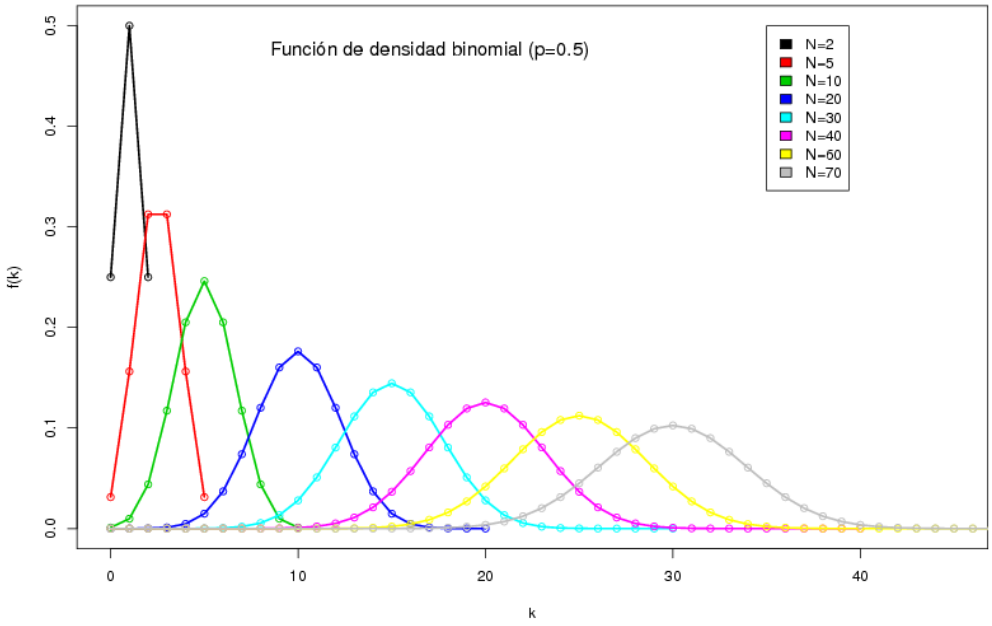


Figura 5.2: Ejemplos de función de densidad binomial (2)

Su función de densidad es:

$$f(k) = \binom{n}{k} p^k q^{n-k} \tag{5.10}$$

donde:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \tag{5.11}$$

la esperanza y la varianza se definen como:

$$E[x] = np \tag{5.12}$$
$$Var[x] = npq \tag{5.13}$$

- Modelo de Poisson

  Es equivalente al modelo binomial cuando p tiende a 0 y n es muy grande, sirve pues para modelizar la probabilidad de que se repitan k veces sucesos improbables a lo largo de un período de tiempo largo. Una de sus primeras aplicaciones fue para modelizar las muertes por coz de caballo en el ejército prusiano en el siglo XIX. En Geografía puede utilizarse para modelizar la probabilidad de que aparezcan episodios extremos, por ejemplo más de dos días de precipitación extrema (superior a 100 mm/24 h p.e.) a lo largo de un año sabiendo cuantas veces ocurre dicho suceso, por término medio, al cabo del año.

  Su función de densidad es:

$$f(k) = \frac{\lambda^k}{k!} e^{-\lambda} \tag{5.14}$$

  y su esperanza y varianza son:

$$E[k] = Var[k] = \lambda \tag{5.15}$$

- Distribución binomial negativa

  La utilización de la distribución binomial o de Poisson para estudiar el número de días de lluvia al mes tiene el inconveniente de asumir que la probabilidad de ocurrencia de lluvia un día es independiente de lo que haya ocurrido el día anterior, evidentemente esto no es así. Para fenómenos en los que se produce agrupación es preferible utilizar la distribución binomial negativa utilizada en ecología para estudiar la distribución espacial de individuos de una misma especie. Su función de densidad es:

$$f(k) == \binom{-r}{k} p^r (-q)^k \tag{5.16}$$

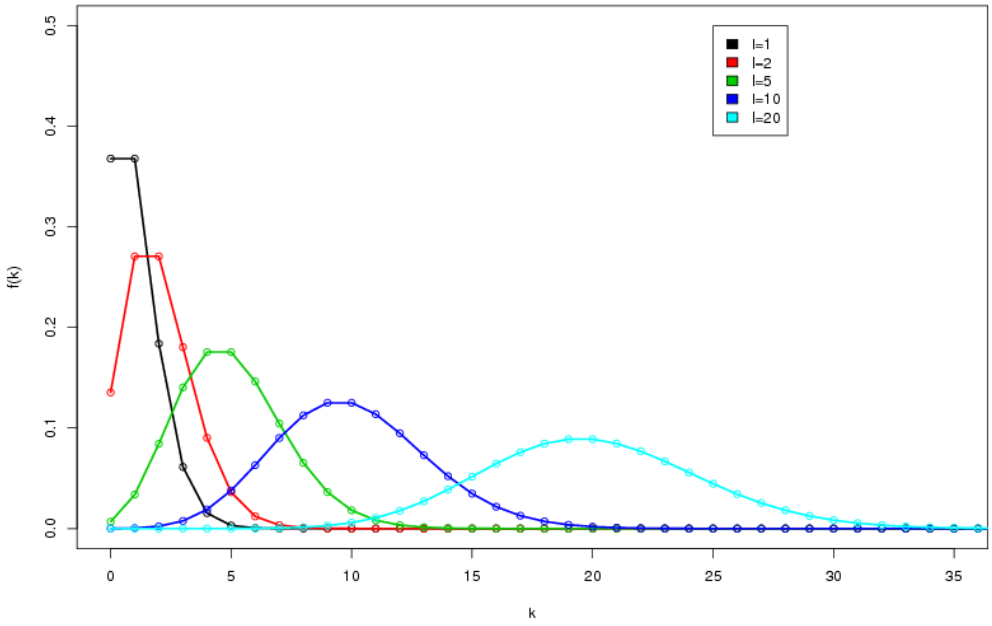


Figura 5.3: Ejemplos de Función de densidad de Poisson

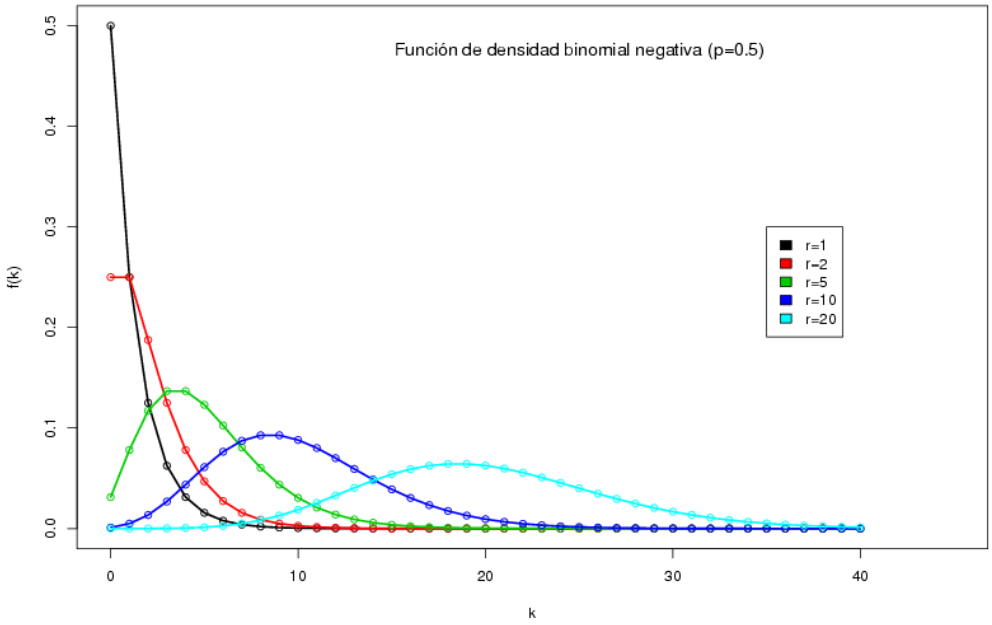


Figura 5.4: Ejemplos de Función de densidad binomial negativa

y su esperanza y varianza son:

$$E[X] = \frac{rq}{p} \tag{5.17}$$

$$Var[X] = \frac{rq}{p^2} \tag{5.18}$$

Representa la probabilidad de que ocurran k fracasos en un experimento antes de que se alcancen r exitos para un suceso de probabilidad p.

Las figuras 5.4 y 5.5 muestran como varía esta función de densidad cuando se mantienen fijos los parámetros p y r respectivamente.

5.2 Modelos de distribución para variables continuas

- Modelo de Gauss (distribución normal).

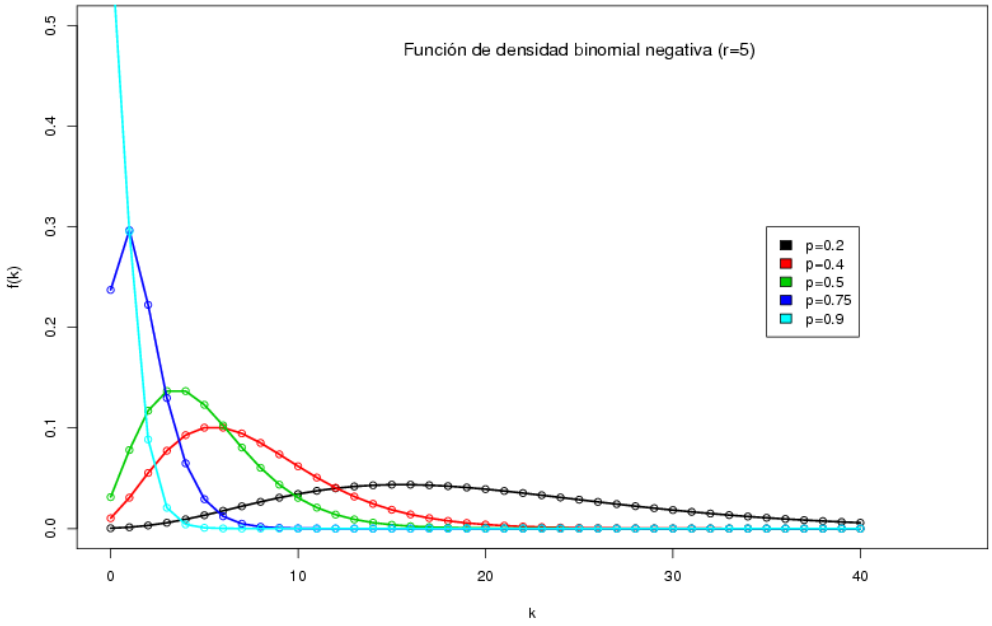


Figura 5.5: Ejemplos de Función de densidad binomial negativa (2)

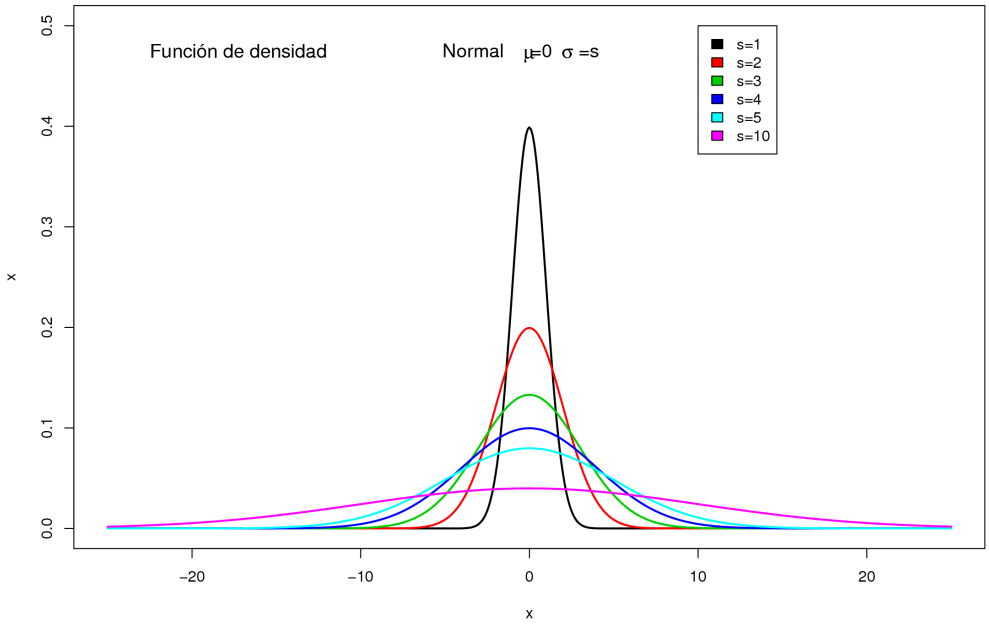


Figura 5.6: Ejemplos de función de densidad normal

Es el modelo más utilizado en estadística, utiliza como parámetros la media (μ) y la desviación típica (σ) de la variable.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} exp - \frac{(x-\mu)^2}{2\sigma^2} \tag{5.19}$$

se cumple que $E(x) = \mu$ y $Var[x] = \sigma^2$

En Geografía Física se suele utilizar con variables que, como la precipitación anual, no alcanzan el cero y muestran histrogramas más o menos simétricos.

La función de distribución normal no puede obtenerse integrando la función de densidad. Puede calcularse sin embargo a partir de métodos numéricos, básicamente la idea consiste en sustituir la integral por un sumatorio y tratarlo como si se tratara de una variable discreta, el resultado será adecuado si dx se hace suficientemente pequeño en un proceso inverso al presentaado en la figura . Puesto que el calculo de F(x) implica el cálculo del área bajo la curva f(x) a la derecha del valor x, si discretizamos la curva en un conjunto de estrechos rectángulos de anchura Δx (figura 5.8) podremos aproximar el valor de ese área como la suma de las áreas de los rectángulos según la ecuación:

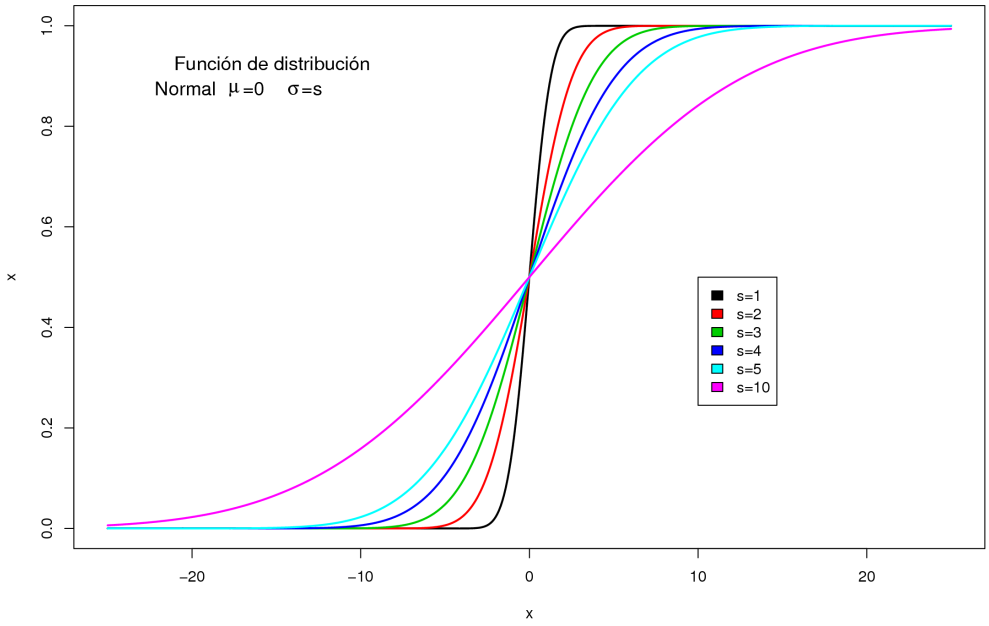


Figura 5.7: Ejemplos de función de distribución normal

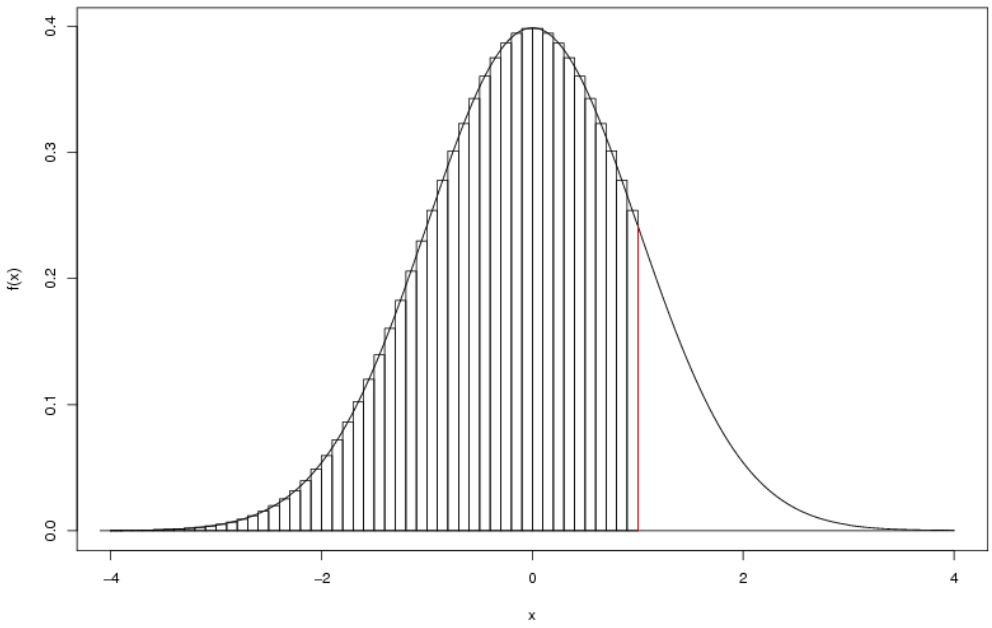


Figura 5.8: Cálculo de la integral de la función de densidad normal con métodos numéricosl

$$F(x) = prob(i \leq x) = \int_{i=-\infty}^{x} f(i)di = \sum_{i=-\infty}^{x} f(i)\Delta i \tag{5.20}$$

La transformación de di en Δi en la ecuación 5.20 es común en los métodos numéricos y se verá más adelante al hablar de modelos físicos.

- Distribución log-normal

  Es la que sigue una variable (x) cuyos logaritmos siguen una distribución normal.

$$y = log(x - a) \tag{5.21}$$

  La inclusión del parámetro a permite trabajar con variables cuyo valor pueda ser inferior a cero, puesto que el logaritmo de un número negativo no está definido. Así $a = min(x)$ si $min(x) < 0$ y $a = 0$ en caso contrario.

$$f(y) = \frac{1}{\sigma_y\sqrt{2\pi}} exp - \frac{(y - \mu_y)^2}{2\sigma_y^2} \tag{5.22}$$

  Se utiliza con variables que tienen una cota inferior (como la precipitación mensual) lo que produce un sesgo positivo.

- Modelo exponencial negativo

  Su función de densidad es:

$$f(x) = \alpha e^{-\alpha x} \tag{5.23}$$

$$E(x) = \frac{1}{\alpha} \tag{5.24}$$

$$Var[x] = \frac{1}{\alpha^2} \tag{5.25}$$

  Suele utilizarse con variables que suelen tener valores pequeños pero que en ocasiones pueden tener valores altos. Por ejemplo se ha utilizado para modelizar la precipitación caida en los diferentes *pulsos*[1] de un episodio de precipitación.

[1] Cuando se trabaja con datos procedentes de pluviógrafo, se llama pulso a cada uno de los intervalos en que se discretiza el tiempo, suelen oscilar entre 10 segundos y 5 minutos

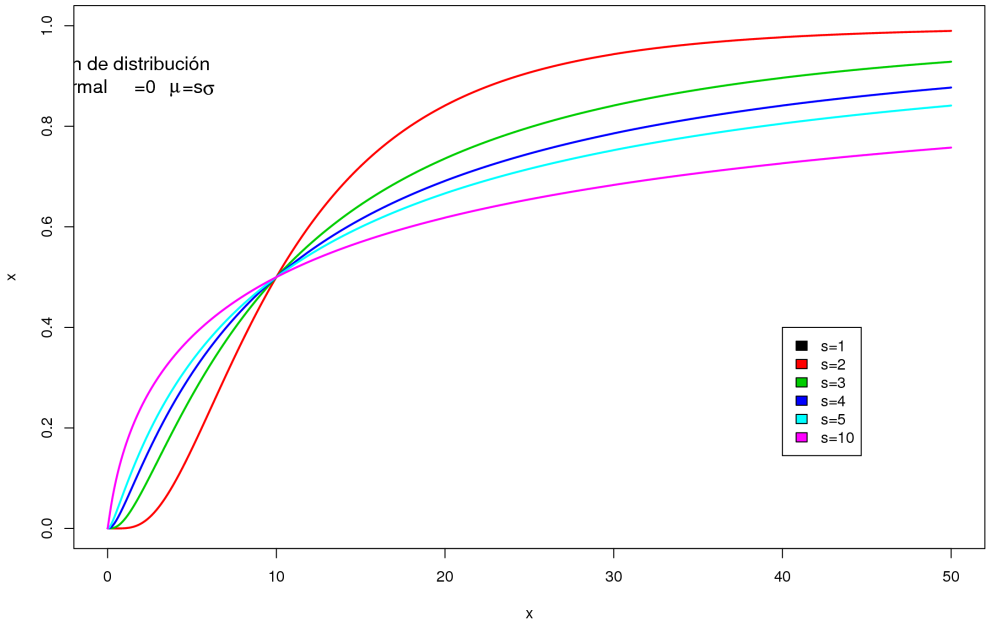


Figura 5.9: Ejemplos de Función de densidad lognormal

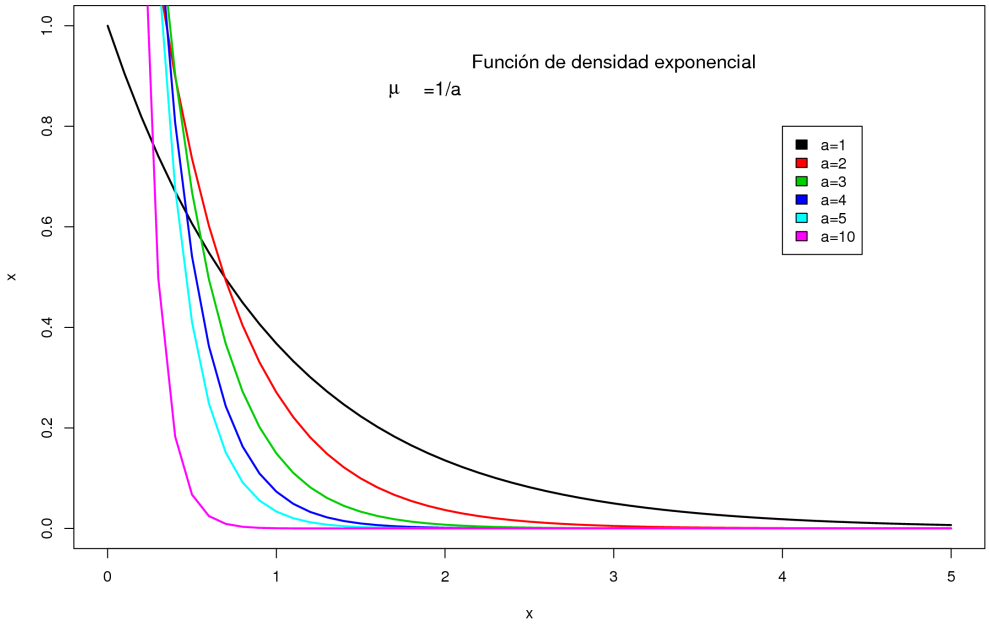


Figura 5.10: Ejemplos de función de densidad exponencial negativa

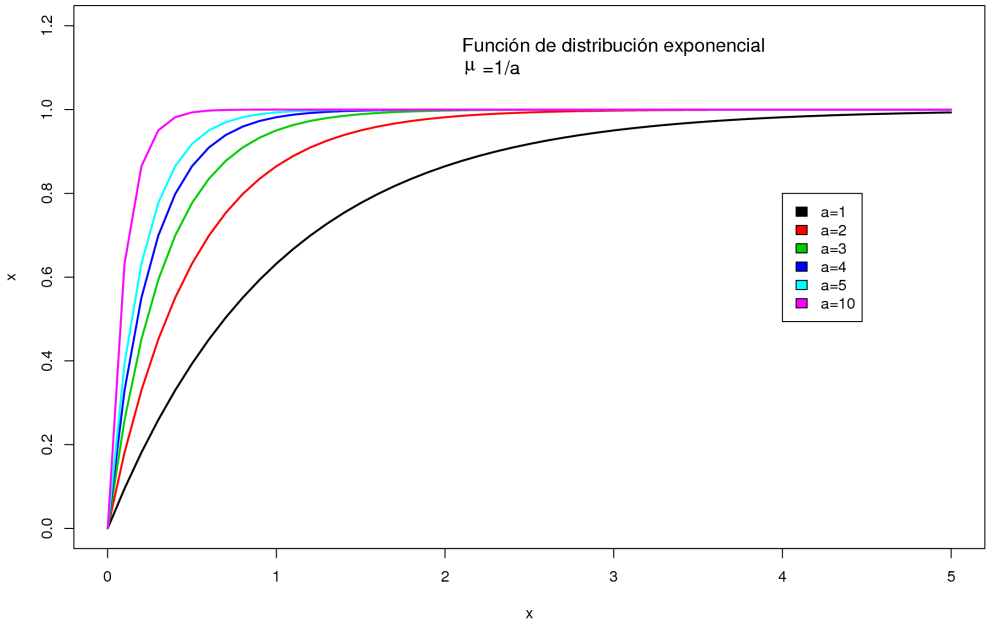


Figura 5.11: Ejemplos de función de distribución exponencial negativa

- Modelo gamma

  Su función de densidad es:

$$f(x) = \frac{x^{a-1}e^{-}x/s}{s^a\Gamma(a)} \tag{5.26}$$

  donde:

$$\Gamma(a) = \int_0^\infty x^{p-1}e^{-x}dx \tag{5.27}$$

  Si a es un número entero se cumple que $\Gamma(a) = (a-1)!$

  El parámetro a es un parámetro de forma y s es un parámetro de escala, ambos deben ser positivos y se cumple que:

$$E[x] = as \tag{5.28}$$
$$Var[x] = as^2 \tag{5.29}$$

  En algunos casos da mejores resultados para la precipitación mensual que la distribución log-normal.

- Modelo Gumbel

$$f(x) = \frac{e^{-\frac{x-u}{\beta}}e^{e^{-\frac{x-u}{\beta}}}}{\beta} \tag{5.30}$$

$$F(x) = e^{-e^{-\frac{x-u}{\beta}}} \tag{5.31}$$

  siendo u un parámetro de localización y β un parámetro de escala. Si te das cuenta, la distribución de Gumbel es una doble exponencial por lo que se utiliza con datos aún más extremos que aquellos para los que se utiliza la exponencial por ejemplo precipitaciones máximas en 24 horas en climas semiáridos.

$$E[x] = u + 0.57721\beta \tag{5.32}$$
$$Var[x] = \frac{\pi^2}{6}\beta^2 \tag{5.33}$$

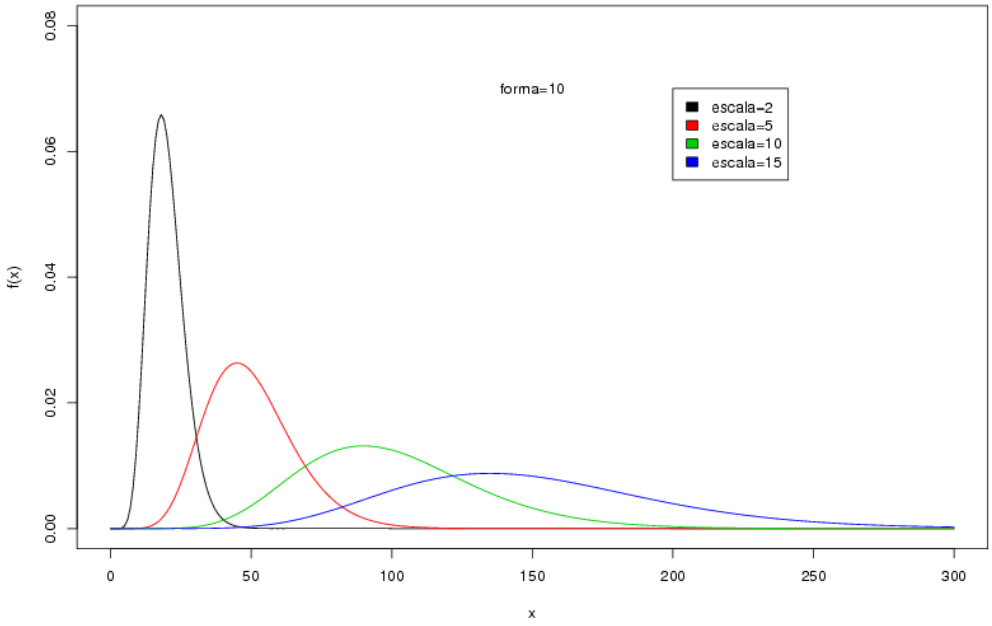


Figura 5.12: Ejemplos de función de densidad gamma

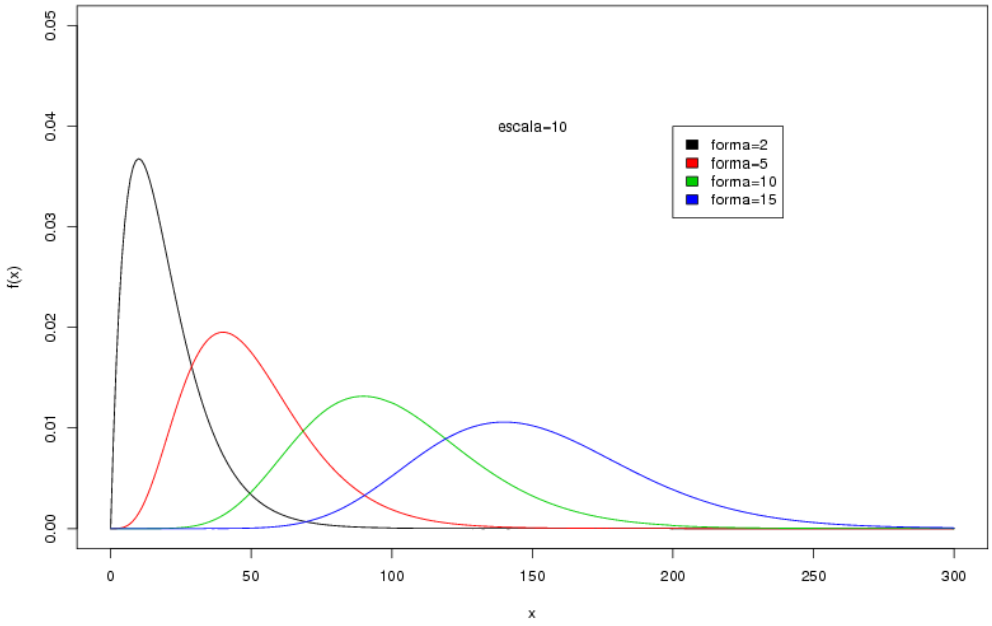


Figura 5.13: Ejemplos de función de densidad gamma

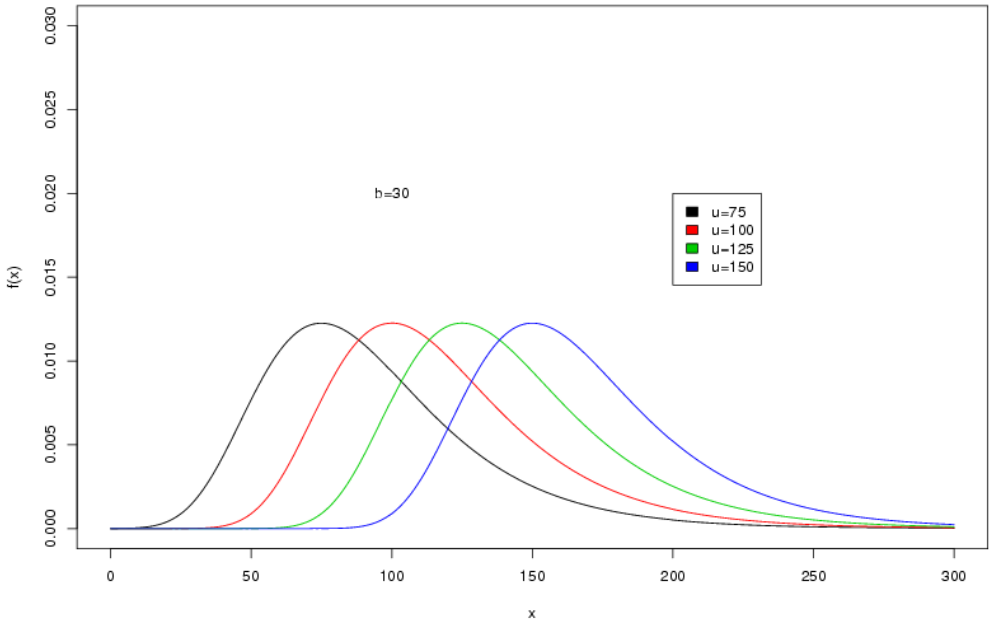


Figura 5.14: Ejemplos de Función de densidad Gumbel

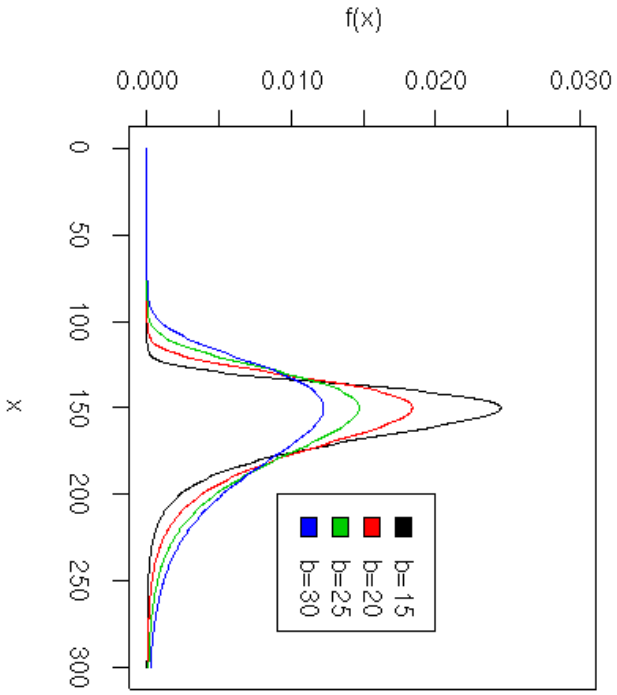


Figura 5.15: Ejemplos de Función de densidad de Gumbel

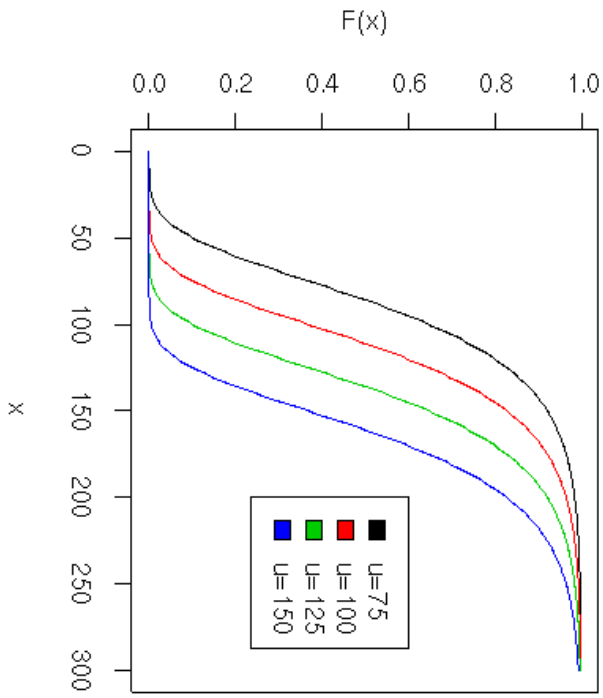


Figura 5.16: Ejemplos de Función de distribución de Gumbel

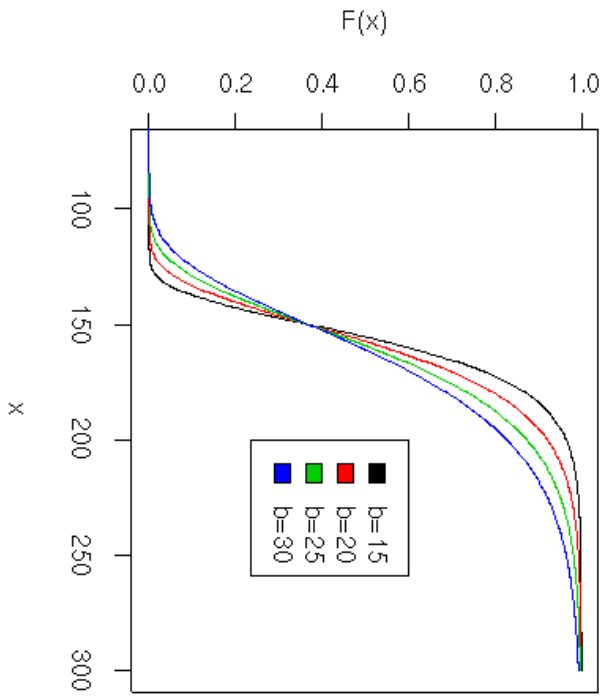


Figura 5.17: Ejemplos de Función de distribución de Gumbel

5.3 Con R

R permite trabajar con muchas distribuciones de probabilidad, y evidentemente con las que hemos visto con anterioridad. Para cada distribución se incluyen cuatro funciones que comienzan con las letras d, p, q y r; a las que se añade un apócope del nombre de la distribución (binom, pois, norm, lnorm, exp y gamma).

- **d** devuelve la función de densidad para el valor x y los parámetros correspondientes.
- **p** devuelve la función de distribución para el valor x y los parámetros correspondientes.
- **r** devuelve n valores extraidos al azar de una población que sigue la distribución de que se trate con los parámetros correspondientes.
- **q** devuelve el valor que tiene un valor de función de distribución F con los parámetros correspondientes para el modelo correspondiente.

- Función binomial

  `dbinom(x,n,p)`

  `pbinom(x,n,p)`

  `qbinom(F,n,p)`

  `rbinom(n,n,p)`

- Función de Poisson

  `dpois(x,l)`

  `ppois(x,l)`

  `qpois(F,l)`

  `rpois(n,l)`

- Función binomial negativa

  `dnbinom(x,n,p)`

  `pnbinom(x,n,p)`

  `qnbinom(F,n,p)`

  `rnbinom(n,n,p)`

- Función normal

  `dnorm(x,m,s)`

  `pnorm(x,m,s)`

  `qnorm(F,m,s)`

  `rnorm(n,m,s)`

- Función lognormal

  ```
  dlnorm(x,m,s)
  plnorm(x,m,s)
  qlnorm(F,m,s)
  rlnorm(n,m,s)
  ```

- Función exponencial

  ```
  dexp(x,l)
  pexp(x,l)
  qexp(F,l)
  rexp(n,l)
  ```

- Función gamma

  ```
  dgamma(x, shape, scale)
  pgamma(x, shape, scale)
  qgamma(F, shape, scale)
  rgamma(n, shape, scale)
  ```

La distribución de Gumbel no está incluida en la distribución básica de R, sin embargo puedes encontrar funciones equivalentes a las anteriores en el fichero `funciones.R`.

```
dgumbel(x, u, b)
pgumbel(x, u, b)
qgumbel(F, u, b)
rgumbel(n, u, b)
```

5.4 ¿Cómo estimar los parámetros de una función de distribución?

Toda función de distribución se basa en un pequeño número de parámetros cuya estimación adecuada resulta muy importante. Básicamente existen 2 tipos de métodos:

1. **Método de los momentos**. Consiste en calcular los parámetros de la función de distribución a partir de los estadísticos de la muestra. En todas los ejemplos de función de distribución expuestos anteriormente se ha incluido la relación entre la esperanza y varianza de la distribución y los parámetros de la misma, para aplicar el método de los momentos basta con despejar el parámetro en la ecuación y hacer $E[x] = m$ y $Var[x] = s^2$. El ejemplo más simple sería el de la distribución normal.

2. **Método de máxima verosimilitud**, más aconsejable aunque más complejo. Consiste en buscar los valores de los parámetros que maximicen la probabilidad de haber obtenido la muestra que tenemos, esta probabilidad se calcula mediante la función de verosimilitud:

$$L(x_1, x_2, ..., x_n, \theta) = \Pi_{i=1}^{n} f(x_i, \theta) \tag{5.34}$$

donde θ representa los parámetros.

Normalmente se suele manejar el logaritmo de la función de verosimilitud, la llamada **función log-verosimil**, ya que suele ser más fácil de manejar.

En R es posible obtener parámetros maximoverosímiles para algunas funciones de distribución[2] con la función `fitdistr` de la librería `MASS`.

Con la función `mle` de la librería `stats4` puede obtenerse una estimación máximoverosimil de parámetros para cualquier función de distribución siempre y cuando dispongamos de la función log-verosimil.

Ejemplo

- Queremos estimar los parámetros de una distribución log-normal para la precipitación de octubre en Zarzadilla de Totana, para ello debemos en primer lugar cargar los datos y hacer una transformación logarítmica. Puesto que no existe el logaritmo de cero y la serie de Zarzadilla de Totana incluye ceros, utilizaremos un valor muy pequeño para el parámetro a $a = 0.01$ que no afectará a los resultados.

```
> datos=read.table("../practicas/base_datos.txt",header=T)
> P=datos$precZ
> lnp=log(P+0.01)
```

Para el método de los momentos, debemos obtener los estadísticos muestrales de la variable transformada:

```
> m=mean(lnp);s=sd(lnp)
> m
[1] 2.475504
> s
[1] 2.901924
```

Para el método de máxima verosimilitud debemos utilizar la variable original pero sumándole el parámetro a.

```
> fitdistr(P+0.01,"lognormal")
meanlog sdlog
```

[2] beta, cauchy, chi-squared, exponential, f, gamma, geometric, log-normal, lognormal, logistic, negative binomial, normal, Poisson, t y weibull

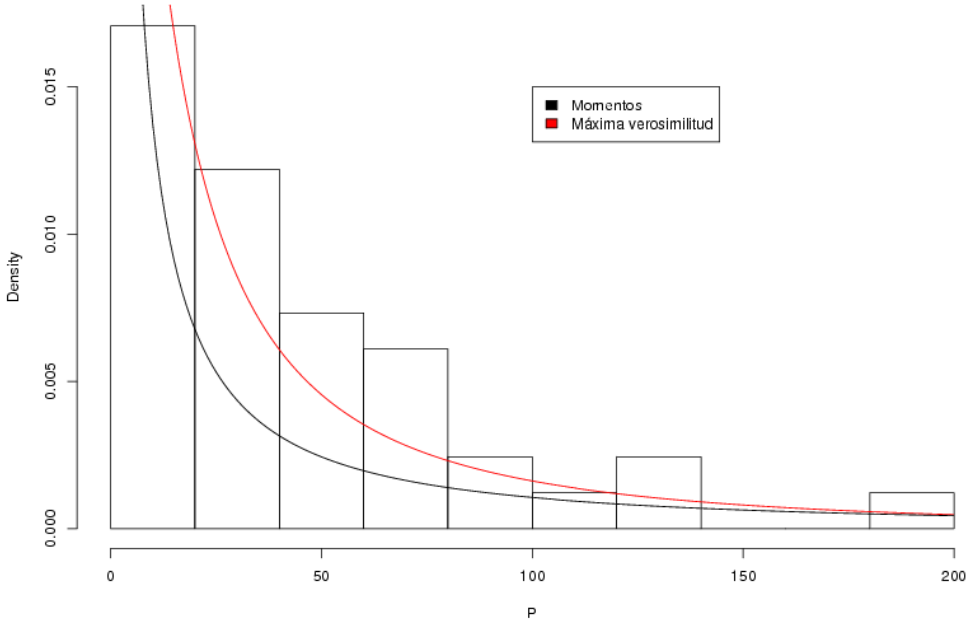


Figura 5.18: Ejemplos de Función de densidad lognormal

```
2.4755045 2.8663164
(0.4476434) (0.3165317)
```

La función fitdistr devuelve, no sólo los valores de los parámetros sino también sus errores típicos. En la figura 5.18 puedes comparar los resultados, sin ser ninguno de los dos especialmente buenos, parece mejor el obtenido con máxima verosimilitud.

5.5 ¿Cual es la distribución que mejor representa mis datos?

Si hemos estimado los parámetros para distintas funciones de distribución y con diferentes parámetros, la pregunta será ahora cual de ellas es la que mejor representa los datos de la muestra, o sea ¿de que tipo de población es más probable que proceda mi muestra?

Existen varios procedimientos para contestar a esta pregunta.

5.5.1 El gráfico Q-Q

Es un método puramente visual, para el que se siguen las siguientes etapas, :

1. Ordena tu muestra (`xs=sort(x)`)
2. Calcula una función de distribución muestral (`Fm=(seq(1:n)-0.5)/n` donde n es el tamaño de la muestra.
3. Puedes representar esta función de distribución muestral para comprobar su aspecto:

   ```
   plot(xs,Fm)
   ```

4. Haz una estimación de los parámetros de la función de distribución que quieres comprobar, si es la normal utiliza la media y varianza muestrales (`m=mean(s);s=sd(x)`).
5. Estima los cuantiles poblacionales a partir de la función de distribución muestral utilizando la función de distribución teórica y los parámetros que quieres comprobar:

   ```
   qp=qnorm(Fm,mean=m,sd=s)
   ```

6. Representa los datos originales ordenados contra los cuantiles estimados y traza una linea recta con pendiente 1 y origen en el cero:

   ```
   plot(xs,qp);abline(0,1)
   ```

   Cuanto más se aproximen los puntos a la linea más cerca estará la muestra a la distribución que has comprobado.

5.5.2 Test χ^2

Consiste en comprarar la frecuencia observada en la muestra correspondiente a determinados intervalos y compararla con la frecuencia esperada si la muestra siguiera una distribución dada. Paso a paso el procedimiento consiste en:

1. Dividir la muestra en m subintervalos, siendo entonces O_j el número de casos observados en el intervalo j y M_j el valor máximo incluido en el intervalo j.
2. Calcular la frecuencia esperada en estos subintervalos como $E_j = (F(M_j) - F(M_{j-1}))N$ donde N es el tamaño muestral.
3. Calcular el estadístico:

$$\chi^2 = \sum_{j=1}^{m} \frac{(O_j - E_j)^2}{E_j} \tag{5.35}$$

4. Si se cumpliera la hipótesis de que la población de procedencia de la muestra sigue exactamente la distribución que se está comprobando, χ^2 seguiriría una distribución chi cuadrado con $m - d - 1$ grados de libertad, donde d es el número de parámetros de la distribución.

El procedimiento para calcular el test χ^2 es bastante tedioso, en el fichero funciones.R tienes algunas funciones para facilitar el proceso.

la función `observados` devuelve el número de casos obtenidos agrupando los elementos de una muestra en intervalos definidos por el usuario, por ejemplo:

```
> x=c(1,2,4,7,9,10,12,18,20,22)
> observados(x,c(0,5,10,15,20,25))
[1] 3 3 1 2 1
```

Si ahora quisiesemos saber cuales serían las probabilidades de esos mismos intervalos si la muestra siguiese una distribución normal con media 2 y desviación típica 5, podríamos utilizar:

```
> espe.norm(m=5,sd=2,co=c(-50,5,10,15,20,50))
[1] 5.000000e-01 4.937903e-01 6.209379e-03 2.866515e-07 3.186340e-14
```

Ten en cuenta la necesidad de ampliar los límites inferior y superior del rango de intervalos para tener en cuenta todos los probablies valores de la distribución comprobada, de manera que la suma de los valores esperados sea 1.

A continuación podemos comparar el número de valores observados en cada intervalo con su correspondiente probabilidad:

```
> o=observados(x,c(0,5,10,15,20,25))
> e=espe.norm(m=5,sd=2,co=c(-50,5,10,15,20,50))
> chisq.test(x=o,p=e)

        Chi-squared test for given probabilities

data:  o
X-squared = 3.138398e+12, df = 4, p-value < 2.2e-16

Warning message:
Chi-squared approximation may be incorrect in: chisq.test(x = o, p = e)
```

El valor de χ^2 es 3.138398e+12 y la probabilidad de obtener un valor igual o superior asumiendo que la población siguiese una distribución normal con media 5 y desviación típica 2 sería $2.2x10^{-16}$ un valor extremadamente bajo que nos llevará a rechazar la hipótesis de que la muestra procede de tal población.

En realidad este método sólo es adecuado cuando el número de casos observados y esperados es mayor que 5 para todas las clases en que se ha dividido la muestra. En caso contrario sería preferible el test de Kolmogorov-Smirnov.

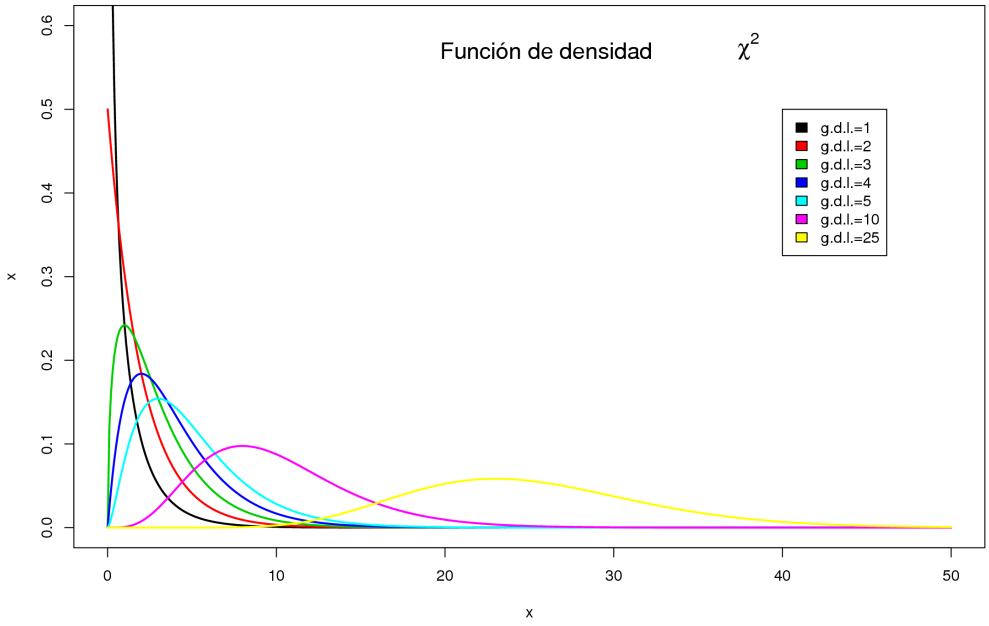


Figura 5.19: Ejemplos de función de densidad χ^2

5.5.3 El test de Kolmogorov-Smirnov

Su aplicación es muy sencilla y es similar al gráfico Q-Q:

1. Ordena tu muestra (`xs=sort(x)`)
2. Calcula una función de distribución muestral (`Fm=(seq(1:n)-0.5)/n` donde n es el tamaño de la muestra.
3. Haz una estimación de los parámetros de la función de distribución que quieres comprobar, si es la normal utiliza la media y varianza muestrales (`m=mean(s);s=sd(x)`).
4. Calcula los valores de la función de distribución teórica para la serie ordenada: `Ft=pnorm(xs,mean=m,sd=s)`
5. El estadístico D de Kolmogorov-Smirnov es el valor máximo de los valores absolutos de las diferencias entre ambas distribuciones: `D=max(abs(Ft-Fm))`

En realidad en R basta con llamar a la función:

```
> ks.test(x, "pnorm", m, s)
```

El resultado dl test será algo parecido a esto:

```
        One-sample Kolmogorov-Smirnov test

data:  x
D = 0.1324, p-value = 0.3162
alternative hypothesis: two-sided
```

donde obtendremos el valor de D y la probabilidad de equivocarnos si admitimos que los datos no proceden de la distribución que queremos comprobar. En este caso deberíamos admitir que los datos proceden de una distribución normal.

Tema 6

Relación entre variables

En el tema anterior se vio como obtener test de significación para la comparación de medias y de muestras, así como para la verificación de la hipótesis de que una muestra proceda de una población que sigue una determinada función de distribución. En este tema se va a estudiar la hipótesis de que dos variables tengan una cierta relación.

Cuando ambas variables son cuantitativas, se hará un análisis de regresión, si una es cualitativa y la otra cuantitativa se hará un análisis de varianza.

6.1 Análisis de regresión

El análisis de regresión permite establecer una relación entre dos variables cuantitativas de manera que podemos predecir una en función de la otra.

Uno de los objetivos más interesantes en las Ciencias Experimentales es la obtención de un modelo matemático que relacione dos o más magnitudes a partir de observaciones experimentales. En la mayor parte de los casos no es posible obtener una relación exacta, sino tan solo una dependencia aproximada que puede medirse numéricamente y a la que denominamos **correlación**. Una vez calculada esta correlación, y asumiendo que puede considerarse estadísticamente significativa, el problema consiste en establecer una relación funcional de tipo $y = f(x)$ que represente de forma aproximada dicha relación, se trata de una **ecuación de regresión** y puede ser de diversos tipos (figura 6.1).

- Lineal $y = ax + b$
- Cuadrática o parabólica $y = ax^2 + bx + c$
- Exponencial $y = ae^{bx}$
- Potencial $y = ax^b$

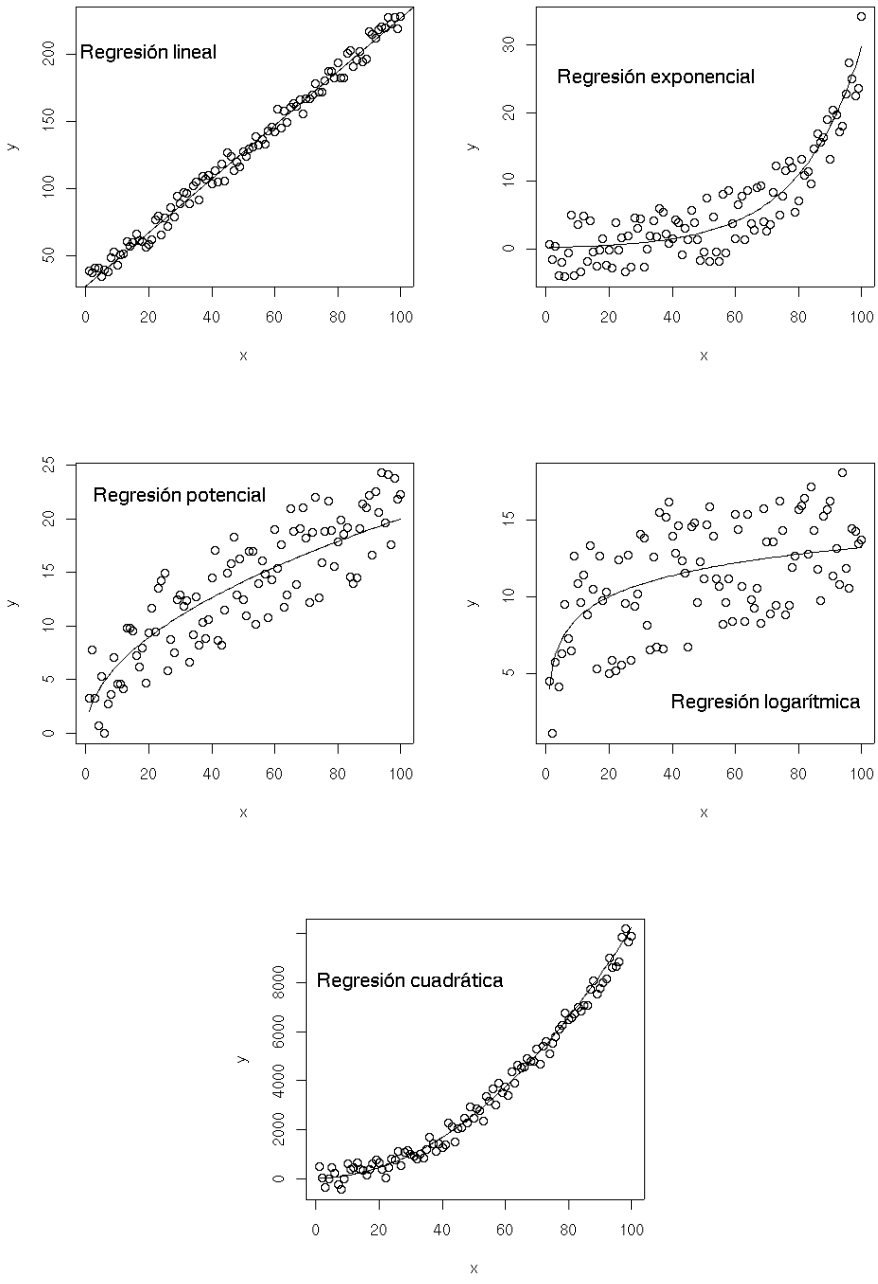


Figura 6.1: Tipos de regresión

- Logarítmica $y = alog(x) + b$

La mejor estrategia para determinar el tipo de relación funcional que se establece entre dos variables es la representación gráfica sobre un sistema de ejes coordenados de los puntos (x_1, y_1),$(x_2, y2)$,...(x_n, y_n) y comprobar la tendencia de los mismos.

Una vez seleccionado un tipo (o varios en caso de duda) el siguiente paso sería calcular sus parámetros. El caso más sencillo es, evidentemente el de la regresión lineal, pero los casos no lineales anteriormente listados pueden transformarse en ecuaciones lineales ($Y = AX + B$) utilizando un simple cambio de variables.

- Lineal: $Y = y$, $A = a$, $B = b$, $X = x$
- Exponencial: $Y = ln(y)$, $A = a$, $B = ln(b)$, $X = x$
- Potencial: $Y = ln(y)$, $A = a$, $B = ln(b)$, $X = ln(x)$
- Logarítmica: $Y = y$, $A = a$, $X = ln(x)$, $B = b$

La regresión cuadŕatica, y en general cualquier regresión polinómica de tipo $y = f(x_1, x_2., , , .x_n)$, se resuelve como un caso particular de la regresión múltiple que se verá posteriormente.

Una vez realizados los cambios de variables adecuados es necesario el calculo de algunos estadísticos básicos:

- Medias $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$
- Varianzas $s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$ $s_y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n}$
- Covarianza $s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}$
- Coeficiente de correlación $r = \frac{s_{xy}}{s_x s_y}$
- Pendiente de la recta que relaciona ambas variables: $A = \frac{s_{xy}}{s_x^2}$
- Ordenada en el origen, es decir el valor que adopta y cuando $x = 0$: $B = \bar{y} - A\bar{x}$

Una vez calculados A y B es necesario deshacer los cambios de variables para obtener a y b a partir de A y B.

6.1.1 Resultados de un análisis de regresión

Un análisis de regresión debe darnos:

- Los valores de los coeficientes de regresión y sus errores típicos
- Los valores de y estimados por el modelo ($\hat{y}$) que son los puntos rojos en la figura 6.2

- Los residuales, o sea la diferencia entre los valores reales de y y los estimados por el modelo ($y - \hat{y}$) que se representan en la figura 6.2 mediante lineas azules cuyas longitudes son los valores de los residuales
- Una medida de la capacidad del modelo para predecir los valores de y, se suele utilizar el llamado coeficiente de determinación (R^2):

$$R^2 = 1 - \frac{\sum(\hat{y_i} - y_i)^2}{\sum(y_i - \bar{y})^2} \tag{6.1}$$

  este índice se interpreta también como el tanto por ciento de la varianza de y debido a la influencia de x.

Puedes observar que el numerador de la ecuación de R^2 es la suma de los cuadrados de las longitudes de las lineas azules del gráfico inferior de la figura 6.2, mientras que el denominador es la suma de los cuadrados de las longitudes de las lineas azules del gráfico superior.

De este modo la hipótesis nula en un análisis de regresión es que $\rho = 0$, $A = 0$ y $B = \bar{y}$ y por tanto la mejor estimación posible de y sería el valor de la media $\hat{y} = \bar{y}$. La hipótesis alternativa sería que $\rho \neq 0$, $A \neq 0$ y $B \neq \bar{y}$ y por tanto la mejor estimación de y se obtendría con la ecuación $\hat{y} = A * x + B$.

6.1.2 Con R

Para pasar de un simple cálculo de correlaciones a un análisis de regresión utilizamos la función `lm`. Es una de las funciones más versátiles y potentes de R y permite calibrar modelos lineales de diverso tipo.

Para ver la forma de ejecutarla vamos a simular dos variables correlacionadas:

```
x=rnorm(100,mean=10,sd=5)
y=x+rnorm(100)
plot (x,y) m=lm(y~x)
```

La orden `plot` se ha introducido para verificar que efectivamente la variable x parece explicar y.

En la llamada a `lm`, `y~x` significa $y = f(x)$. La variable m contiene ahora el modelo del que podemos obtener diversa información aplicandole las funciones adecuadas:

- `summary(m)`

```
Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-2.0813 -0.6262  0.1284  0.6356  2.0113

Coefficients:
```

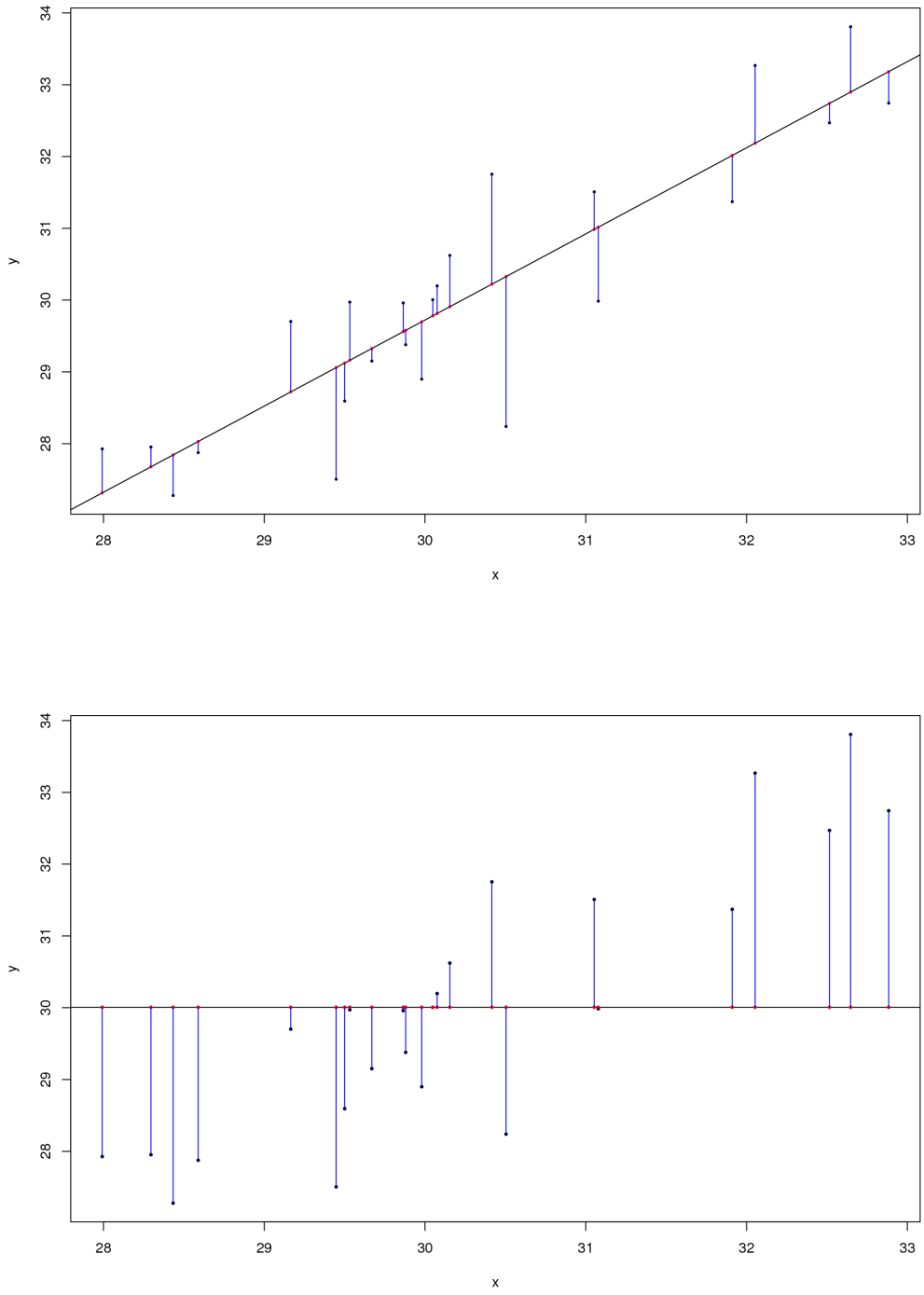


Figura 6.2: Residules de un modelo basado en la media y otro basado en regresión lineal

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.15529    0.21212   0.732    0.466
x            0.96784    0.02066  46.850   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.913 on 98 degrees of freedom
Multiple R-Squared: 0.9573,     Adjusted R-squared: 0.9568
F-statistic:  2195 on 1 and 98 DF,  p-value: < 2.2e-16
```

Nos devuelve el mínimo y el máximo de los residuales junto a los valores de q_{25}, q_{50} y q_{75}; las estimaciones de los coeficientes A y B (en este caso $A = 0.155$ y $B = 0.968$ junto con sus errores típicos y un indicador de la potencia de la variable x como estimador de y que son los valores de `Pr(>|t|)`, cuanto menores sean y más estrellas aparezcan a su lado mejor; el valor de r^2, en este caso 0.9573 y finalmente un estadístico F cuyo p-value asociado nos indica la probabilidad de obtener por azar una regresión tan buena o más que la que hemos obtenido sui se cumple la hipótesis nula.

- `fitted(m)` nos dará los valores de y predichos por el modelo ($\hat{y}$)
- `residuals(m)` la diferencia entre el valor real de y y el predicho por el modelo ($e = y - \hat{y}$)

Ejercicios

- Calcula tres modelos que permitan estimar la precipitación de octubre en Zarzadilla de Totana a partir de la de Librilla, El Algar y Yecla. ¿Cual de ellos daría mejores resultados?

6.2 Regresión y correlación múltiple

Es una extensión del caso de dos variables, sean x_1, x_2, ..., x_n variables aleatorias con medias μ_1, μ_2, ..., μ_n y varianzas σ_1^2, σ_2^2, ..., σ_n^2. El objetivo será hallar unos coeficientes b_i (i=1,2,...,n) tales que:

$$\hat{x_n} = b_0 + b_1 x_1 + b_2 x_2 + ... + b_{n-1} x_{n-1} \tag{6.2}$$

se ajuste a x_n según el criterio de los mínimos cuadrados.

Los valores de b_i se obtienen a partir de las ecuaciones $b_i = \beta_i \frac{\sigma_n}{\sigma_i}$ y los valores de β_i se obtienen resolviendo el sistema:

$$\beta_1 + r_{1,2}\beta_2 + ... + r_{1,m-1}\beta_{n-1} = r_{1,n} \tag{6.3}$$
$$r_{1,2}\beta_1 + \beta_2 + ... + r_{2,m-1}\beta n - 1 = r_{1,n} \tag{6.4}$$
$$..... \tag{6.5}$$
$$r_{1,n-1}\beta_1 + r_{2,n-1}\beta_2 + ... + \beta_{n-1} = r_{n-1,n} \tag{6.6}$$

En el caso de la regresión polinómica, por ejemplo parabólica, simplemente habrá que hacer $x_1 = x$, $x_2 = x^2$, ..., $x_n = x^n$.

6.2.1 Con R

Para llevar a cabo un análisis de regresión múltiple con R basta con añadir más variables independientes a la ecuación del modelo:

```
m=lm(y~x_1+x_2+...+x_n)
```

e interpretar los resultados de forma similar a como se hace con la regresión simple.

Ejercicios

- Haz un modelo que estime la precipitación total de octubre en Zarzadilla de Totana a partir de los datos de precipitación máxima diaria y el número de días de precipitación.
- ¿Cual de las dos variables es la que tiene más peso en el modelo?

6.3 Análisis de varianza

El análisis de varianza es adecuado cuando la variable independiente es de tipo cualitativo, en este caso se denomina **factor** y tiene uno o varios **niveles** o clases; nos ayuda a determinar si la pertenencia a una clase influye de forma significativa en el valor de la variable cualitativa.

En la práctica se trata de comparar las medias de cada una de las clases asumiendo, aunque pueda parecer contradictorio con el nombre del análisis, que las varianzas son iguales y que las distribuciones son normales.

Si hay un sólo factor con dos niveles se utiliza la t de Student, con 3 o más niveles el *one-way Anova*. En caso de que haya varios factores se utilizará el *two-way* o *three-way Anova*. Si toda combinación de niveles tiene una replicación se denomina *Diseño factorial* y permite analizar todas las posibles interacciones entre los factores.

Hay que tener en cuenta que los términos independiente y dependiente no hacen necesariamente referencia a causa y efecto respectivamente. Así la pregunta ¿Existe relación entre los usos del suelo y la elevación? puede contestarse con un análisis de varianza en el que el factor son los usos del suelo y la variable independiente la elevación. Pero es evidente que es la elevación la que, en todo caso, condicionará los usos del suelo.

En la figura 6.3 se muestra un ejemplo hipotético en el que se representa una variable x y un factor V con tres niveles representados por los colores rojo, verde y azul. La linea horizontal representa la media.

La varianza total de la muestra puede calcularse como:

$$SS_t = \sum_{i=1}^{n} (x_i - m_x)^2 \tag{6.7}$$

la varianza debida al factor como:

$$SS_f = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (m_j - m_x)^2 \tag{6.8}$$

y la varianza de error como:

$$SS_e = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x - m_j)^2 \tag{6.9}$$

cumpliendose que:

$SS_t = SS_f + SS_e$

Es decir que la variación general equivale a la variación debida al factor analizado más una variación residual que puede ser totalmente aleatoria o incluir variación debida a otros factores.

En el ejemplo que nos ocupa:

$SS_t = s_x^2 N = 105.26$,
$SS_m = 94.09$ y
$SS_e = 11.173$

La cuestión es si podemos considerar que el factor V explica un porcentaje de la variación suficientemente alto como para rechazar la hipótesis nula de que no existe relación para un nivel de significación determinado.

Para ello debemos calcular un estadístico

$$F = \frac{\frac{SS_f}{k-1}}{\frac{SS_e}{n-k}} \tag{6.10}$$

donde n es el tamaño de la muestra y k el número de niveles del factor.

Este estadístico se distribuye, para el caso de que se cumpla la hipótesis nula, según una función F de Snedecor con k-1 y n-k grados de libertad. Por tanto deberemos determinar cual es la probabilidad de obtener el valor del estadístico F obtenido o superior.

Como puedes ver, de forma similar a lo que ocurría con el análisis de regresión, el numerador de la ecuación de es la suma de las lineas azules del gráfico inferior de la figura 6.3 al cuadrado, mientras que el denominador es la suma de las lineas azules del gráfico superior al cuadrado.

En nuestro caso F es igual a 75.79, y la probabilidad buscada es prácticamente 0 por lo que podemos rechazar la hipótesis nula sin ningún problema.

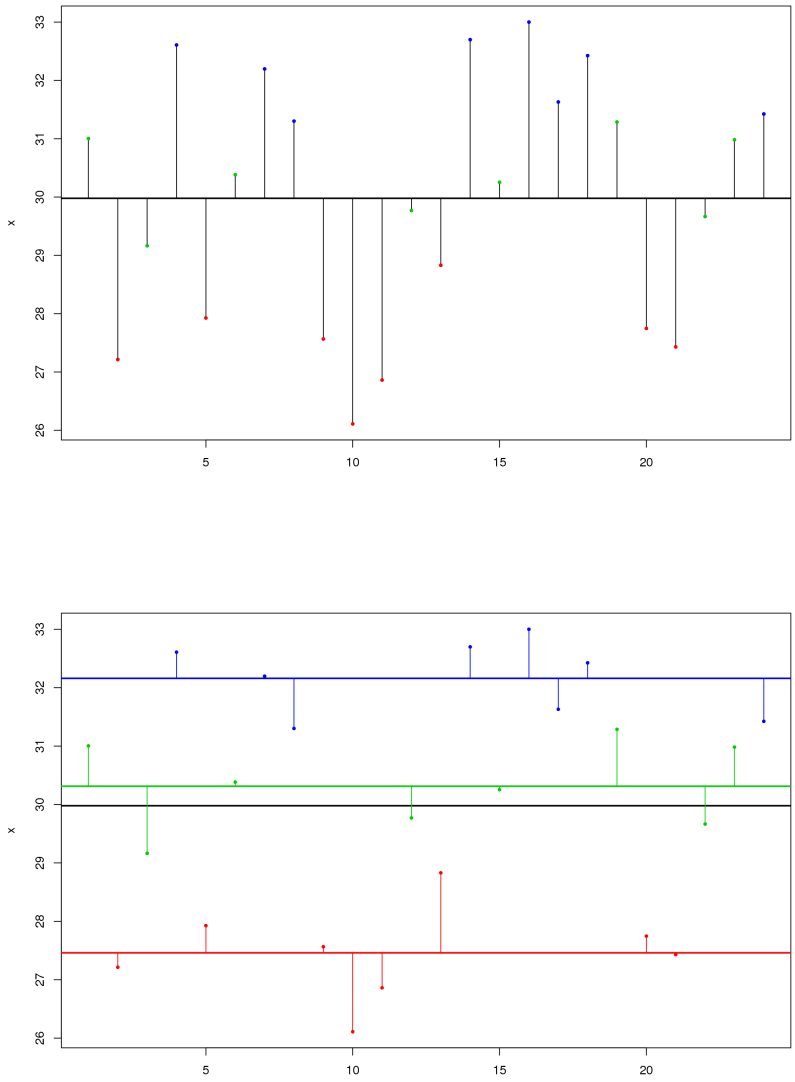


Figura 6.3: Residules de un modelo basado en la media y otro basado en el análisis de varianza

6.3.1 Con R

Puedes obtener la probabilidad de igualar o superar un determinado valor de F mediante:

```
1-pf(F,k-1,n-k)
```

aunque puedes hacer el análisis de varianza directamente con la orden:

```
modelo=lm(x~v)
```

Si v fuera un factor codificado mediante números habría que utilizar:

```
modelo=lm(x~as.factor(v))
```

para especificar que v debe considerarse como un factor y no como una variable cuantitativa.

Si ahora tecleamos:

```
summary(modelo)
```

tendremos la siguiente respuesta:

```
Call:
lm(formula = x ~ as.factor(v))

Residuals:
     Min       1Q   Median       3Q      Max
-1.35058 -0.55770  0.05306  0.48381  1.36946

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    27.4606     0.2579 106.481  < 2e-16 ***
as.factor(v)2   2.8530     0.3647   7.823 1.18e-07 ***
as.factor(v)3   4.6999     0.3647  12.886 1.93e-11 ***
---
Signif. codes:  0 ’ ***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 0.7294 on 21 degrees of freedom
Multiple R-Squared: 0.8892,     Adjusted R-squared: 0.8787
F-statistic:  84.3 on 2 and 21 DF,  p-value: 9.248e-11
```

que incluye los siguientes elementos más significativos:

- La fórmula con que se especificó el modelo
- Una caracterización estadística de los residuales
- Los coeficientes, cuyo valor estimado equivale a la media de X para cada uno de los niveles de V, estos coeficientes deben interpretarse así:

  - $m_x|v = 1 = 27.46$
  - $m_x|v = 2 = 27.46 + 2.853 = 30.3136$
  - $m_x|v = 3 = 27.46 + 4.699 = 32.1605$

  Junto a los coeficientes aparecen sus errores típicos y los valores de `Pr(>|t|` con sus correspondientes estrellas. En general cuanto menor sea el error típico menor será el intervalo de confianza de esas medias y mejor el modelo.

- Un valor de R^2 (parte de la varianza de x explicada por el modelo) y un valor de F y su probabilidad asociada

6.3.2 Ejemplos

- Vamos a comprobar si existen diferencias en al precipitación media entre Zarzadilla de Totana, Librilla y Yecla.

  1. `datos=read.table("base_datos.txt",header=T)`
  2. `Z=datos$precZ;Y=datos$precY;L=datos$precL`
  3. `prec=c(Z,L,Y)`
  4. `obs=factor(c(rep("Z",length(Z)),rep("L",length(L)),rep("Y",length(Y))))`
  5. `modelo=lm(prec obs)`
  6. `summary(modelo)`

  El resultado será:

```
Call:
lm(formula = prec ~ obs)

Residuals:
    Min      1Q  Median      3Q     Max
-48.593 -33.510  -9.412  24.488 177.607

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   48.593      7.116   6.829 3.76e-10 ***
obsY          -5.180     10.064  -0.515    0.608
obsZ          -3.985     10.064  -0.396    0.693
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 45.57 on 120 degrees of freedom
Multiple R-Squared: 0.002416,   Adjusted R-squared: -0.01421
F-statistic: 0.1453 on 2 and 120 DF,  p-value: 0.8649
```

- En muchos modelos distribuidos o semidistribuidos disponemos de capas espaciales de una variable cualitativa pero necesitamos para el modelo una variable cualitativa. Por ejemplo, para un modelo hidrológico necesitamos obtener un mapa de la capacidad de infiltración del suelo, si disponemos de un mapa de suelos y asumimos que la capacidad de infiltración varía con el tipo de suelo, podemos medir la variable en distintos tipos de suelo para:
  - Determinar si se cumple la hipótesis de que la capacidad de infiltración depende, al menos en parte, del tipo de suelo
  - Obtener la capacidad de infiltración media para cada tipo de suelo
  - Reclasificar el mapa original y asignar a cada clase de suelo el valor de su capacidad de infiltración media.

  Este procedimiento se denomina interpolación por clasificación.

Ejercicios

- Interpreta los resultados del análisis de varianza del primer ejemplo.
- Haz un análisis similar para la precipitación máxima diaria

Tema 7

Programación

La utilización de modelos matemáticos implica, en mayor o menor medida, escribir un programa en un lenguaje de programación que implemente el modelo en cuestión. Se trata de una tarea cuya dificultad dependerá de la complejidad del modelo pero que, lógicamente, requiere una fase inicial de aprendizaje.

En R disponemos de múltiples funciones para realizar modelos empíricos, pero no para otro tipo de modelos. Sin embargo R si que dispone de un lenguaje (el lenguaje S) que permite programarlos.

Existen múltiples lenguajes de programación, los ejemplos que van a verse en este curso se han desarrollado lógicamente en lenguaje S para su ejecución con el programa R.

7.1 Introducción a la computación

Un ordenador (*computador*) puede definirse como una máquina programable para el tratamiento de información, es decir, una máquina para realizar **cómputos** (determinación indirecta de una cantidad mediante el cálculo a partir de ciertos datos). Un programa es la descripción detallada de los **algoritmos** necesarios para llevar a cabo un computo en un lenguaje, intermedio entre el lenguaje humano y el lenguaje de la máquina, denominado **lenguaje de programación**.

Un **algoritmo** es un método general de resolución de todos los problemas del mismo tipo en el que se detalla paso a paso lo que debe hacerse. Cada uno de estos pasos costituye una **sentencia** o **instrucción**. Una **instrucción** es una combinación de palabras, variables, constantes y símbolos que, obedeciendo a la sintaxis propia del lenguaje de programación utilizado, indican al ordenador la acción que debe ejecutar.

Un **programa** es de este modo la expresión de un algoritmo en un lenguaje de programación entendible tanto por el ordenador como por los humanos.

A la hora de realizar algún tipo de cómputo en un ordenador, lo normal es utilizar un programa ya desarrollado (una **aplicación**) de propósito general. El problema surge cuando estas aplicaciones no hacen exactamente lo que queremos que hagan. En ese caso la solución puede ser programar el ordenador para que realice la tarea.

De hecho la mayor parte de las aplicaciones informáticas incluyen algún tipo de lenguaje de programación para aumentar su flexibilidad.

7.2 Lenguajes de programación

Desde los años cincuenta se han desarrollado multiples lenguajes de programación. Los lenguajes de **primera generación** eran los ensambladores, lenguajes de bajo nivel que simplemente renombran instrucciones de código máquina, es decir combinaciones de unos y ceros, en un formato más legible por humanos.

Una **segunda generación** esta constituida por lenguajes imperativos en los que la unidad de trabajo del programa es la instrucción (Fortran, Cobol o Basic). Cada instrucción representa varias instrucciones en lenguaje máquina. La evolución de estos da lugar a lenguajes imperativos de **tercera generación** (Modula-2, Algol, Pascal, C).

En una **cuarta generación** aparecen lenguajes aún de más alto nivel orientados a aplicaciones concretas, como SQL para el acceso a bases de datos. Las instrucciones que manejan estos lenguajes manipulan objetos definidos por el usuario (tablas, objetos con diferentes propiedades), no se basan en instrucciones que manejen elementos del *hardware* (CPU, RAM, disco duro), por lo que están más cerca de los conceptos humanos que del lenguaje máquina.

En la **quinta generación** aparecen lenguajes orientados al procesamiewnto de lenguaje natural (Perl o AWK) y lenguajes orientados a objetos (C++ o Java). En estos últimos lo importante no es la instrucción sino el objeto definido como un tipo de dato complejo que intengra tanto cun conjunto de variables como las funciones que pueden ejecutarse con ellas, el lenguaje S pertenece a este tipo. LA unión de lenguajes orientados a procesamiento de lenguaje natural con los programas orientados a objetos dan lugar a nuevos lenguajes muy potentes, aunque poco utilizados, como Python o Ruby.

Derivados de los lenguajes orientados a objetos, aparecen los lenguajes visuales, capaces de generar interfaces gráficas de usuario manejando objetos gráficos como ventanas o botones. Lenguajes como Visual-Basic, Visual-C o Tcl-Tk pertenecen a esta categoría.

La cada vez mayor relevancia de Internet a supuesto l aparición de lenguajes especilmente concebidos para crear páginas web e interpretar las interacciones del usuario con esas páginas web (PHP, por ejemplo).

Otra diferencia importante es entre lenguajes compilados e interpretados. En los primeros se sigue toda la secuencia de fases presentada anteriormente (C, Pascal, Fortran, Cobol). En los segundos no hay compilación ni enlace, el programa es leido por un **intérprete** que lee cada orden, la traduce a lenguaje máquina y la interpreta. Lógicamente, debido a este proceso los programas son más lentos y requieren más recursos del ordenador. La ventaja es que todo el proceso de compilación se simplifica y se permite la portabilidad (los programas pueden ser utilizados bajo diversos sistemas operativos con escasas modificaciones. Ejemplos de estos lenguajes son Basic, Awk, Tcl-Tk o Python.

Aunque existen varios lenguajes en realidad todos responden más o menos a las mismas ideas y elementos básicos. A partir de aquí se utilizará para desarrollar los ejemplos y hacer las prácticas el lenguaje de programación **S** que puede utilizarse para programar aplicaciones dentro del programa **R**. La ventaja fundamental de utilizar

S como lenguaje de programación es que es un lenguaje de altísimo nivel que incluye funciones para la entrada, salida y representación de datos que exigirían muchas lineas de código en otros lenguajes de programación. Por otro lado al ser un lenguaje orientado al análisis de datos, resulta muy adecuado para trabajar en modelización.

La desventaja de R es que, precisamente por ser un lenguaje de muy alto nivel, se vuelve muy lento cuando el programa se complica o cuando el volumen de datos implicado es muy alto. No es un buen lenguaje para hacer grandes programas, pero resulta muy adecuado para pequeñas aplicaciones y modelos.

7.3 Fases en el desarrollo de un programa

Existenten una serie de etapas desde el reconocimiento de la existencia de un problema, susceptible de ser resuelto por un programa hecho por nosotros, y la solución de aquel mediante la ejecución del programa:

- **Análisis del problema**
- **Desarrollo del algoritmo** mediante esquemas o pseudocódigo.
- **Codificación** en un lenguaje de programación concreto.
- **Edición**, transcripción del código al ordenador (código fuente).
- **Prueba de ejecución** y corrección de los errores de ejecución (el programa no se ejecuta correctamente) o de lógica (el programa produce resultados erroneos).
- **Utilización**

Este esquema es válido para la programación con lenguajes interpretados (que suele ser lo habitual en modelización a pequeña escala) como puede ser BASIC, awk, python, Perl o el lenguaje S. En caso de que se trabaje con un lenguaje compilado (C, C++, Java, etc.) habría que introducir antes de la ejecución las fases de:

- **Compilación** o traducción del programa a código máquina y
- **Enlazado**, es decir la inclusión dentro del programa de funciones ya compiladas y que están disponibles en otros ficheros.

7.4 Elementos de programación

7.4.1 Variables y operadores

Todos los programas manipulan datos que se guardan en una o varias posiciones de memoria. Como, evidentemente, sería muy complejo hacer referencia directa a las posiciones de memoria, se les asignan variables cuyo nombre es más sencillo e intuitivo.

Por ejemplo si queremos que una variable contenga el día del mes, podemos escribir en el programa:

```
> dia=3
> dia
[1] 3
```

El ordenador interpretará que queremos reservar una posición de memoria que contenga el número 3 y a la que nos referiremos durante el desarrollo del programa como `dia`.

Si por el contrario escribimos:

```
> x=rep(0,10)
> x
[1] 0 0 0 0 0 0 0 0 0 0
```

el ordenador interpretará que queremos reservar diez posiciones de memoria consecutivas para tratarlas como un vector y que deben contener el número 0.

Las variables pueden manipularse mediante operadores aritméticos o lógicos $(+, -, *, /, \&\&, ||, =, != , >, <, >=, <=)$ que combinan dos variables para producir un resultado que se almacena en otra variable $a = b + c$. También es posible modificar variables:

```
> a=3
> a=a+5
> a
[1] 8
```

Otra posibilidad es almacenar cadenas de caracteres en una variable. Para indicarle al ordenador que toda la cadena es un sólo objeto, esta debe entrecomillarse:

```
> a="Hola"
```

7.4.2 Entrada y salida

La entrada y salida de datos en R se ha tratado ya en el apartado 2.3

7.4.3 Estructuras de control: Bucles

En la gran mayoría de los programas, el proceso de los datos requiere ejecutar alguna tarea repetidas veces o decidir realizar una tarea u otra en función de los valores de alguna variable. Las instrucciones para llevar a cabo estas acciones constituyen las estructuras de control.

Todas las estructuras de control comienzan con una palabra clave como `for`, `if` o `while`.

Existen dos tipos fundamentales de bucles, los bucles *FOR* y *WHILE*, aunque en realidad son intercambiables.

Los bucles FOR en R tienen la siguiente forma:

```
for (i in A){cat(sprintf("Número %d\n",i) }
```

donde `A` es un conjunto de objetos que puede construirse de diversos modos, por ejemplo:

- `A=1:10`
- `A=c(1,2,43,2,1,2)`
- `B=seq(1,10,2)`
- `C=rep(0,10)`
- `A=c(B,C)`
- `A=list(B,C)`

comprueba los resultados que producen los anteriores ejemplos:

La función `sprintf` crea una cadena de texto a partir de una cadena enctrecomillada en la que los %g y %s se sustituyen por las variables que la siguen. Si el código es %g se espera un número y si es %s una cadena de texto. Por ejemplo:

```
 > x=2.34
 > v="mi número"
 > sprintf("El valor de \%s es \%g",v,x)
[1] "El valor de mi número es 2.34"
```

La ejecución del bucle FOR asigna a la variable i los diferentes valores contenidos en el objeto A y ejecutará la orden entre llaves. Por ejemplo:

```
A=seq(1,10)
for (i in A){
      cat(sprintf("%d^2=%d\n",i,i^2))
}
```

devolverá el resultado de elevar al cuadrado los números del 1 al 10. El código \n indica salto de linea.

Un bucle WHILE hace los mismo, sólo que ahora la variable contador se inicializa y se incrementa fuera de la orden. El bucle se ejecuta hasta que deja de cumplirse la condición establecida en él. La siguiente orden nos devolvera los números desde el 1 hasta el primer número cuyo cubo sea mayor que 10000.

```
> i=0;r=0
while (r<10000){
      i=i+1;r=i^3;cat(sprintf("%d^3=%d\n”,i,r))
}
```

Si ejecutas este código comprobarás que el último resultado devuelte r es mayor que 1000. ¿Sabrías decir por qué?

Las instrucciones `break`, `continue` y `exit` alteran también el orden de ejecución.

La primera fuerza al flujo del programa a salir de un bucle `for` o `while` y constinuar con la siguiente orden. La segunda vuelve a la orden inicial del bucle (e incrementa el contador en el caso del bucle FOR). La última sale totalmente del programa.

En este tipo de bucles se suele distinguir entre la **variable contador**, destinada a contener diferentes valores que se van incrementando o decrementando cada vez que el ordenador ejecuta la orden FOR y las **variables acumuladoras** que van a permitir almacenar valores que aumentan o disminuyen de forma no constante durante el proceso. El siguiente ejemplo permite calcular el factorial de un número (variable **n**) que se guardará en la variable acumuladora **fac**:

```
n=6

fac=1

for (i in 1:n){ fac = fac * i }
```

Los bucles son un elemento fundamental en cualquier lenguaje de programación y son muy útiles, sin embargo la presencia en R de muchos bucles largos insertados unos dentro de otros tiende a hacer muy lento el programa. R dispone de métodos alternativos más eficientes como `tapply`.

7.4.4 Estructuras de contro: Toma de decisiones

En ocasiones se debe romper el flujo de un programa y ejecutar un grupo de instrucciones u otras en función de los valores que adopta una variable.

El esquema básico de las instrucciones **if ... else** es:

```
if (condición) {instrucciones1} else {instrucciones2}
```

Donde instrucciones1 es el conjunto de ordenes que se ejecutan si condición se cumple e instrucciones2 las que se ejecutan en caso contrario. Por ejemplo:

```
if (x==3){y=27}else{y=9}
```

Si x es igual a 3 y será igual a 27, sino y será 9.

```
if (a>20){ printf("%d es mayor que 20\n",a)}
else
{printf("%d no es mayor que 20\n",a)}
```

en este caso la respuesta del programa será decirnos si a es o no mayor que 20.

Si en lugar de estar formadas por una sóla instrucción, **instrucciones1** e **instrucciones2** estan compuestas por varias, es necesario construir un bloque con ellas poniendo corchetes a su alrededor, por ejemplo:

```
if (x==3) {y=27;z=3} else {y=9;z=4}
```

7.4.5 Funciones definidas por el usuario

Uno de los elementos básicos de cualquier lenguaje de programación son las funciones. Su objetivo es encapsular una serie de instrucciones que tienen consistencia por si mismas como *subprograma* y que son llamadas de forma habitual y siempre de la misma forma por el programa. Las funciones generan uno o más datos de salida a partir de unos datos de entrada.

Existen múltiples funciones ya disponibles en cualquier lenguaje de programación, en concreto el lenguaje S se caracteriza por un gran número de funciones de análisis de datos que ya has visto, sin embargo puede ocurrir que necesitemos una función no disponible, en ese caso podemos definirla y utilizarla en cualquier momento dentro de nuestro programa.

Cuando un programa se va haciendo más sofisticado, necesita llevar a cabo determinadas acciones varias veces. Pero no resulta una técnica eficiente de programación el repetir varias veces el mismo conjunto de órdenes.

En su lugar es preferible incluir estas órdenes como una *función definida por el usuario* con ello se consigue:

- Dividir el código en un número pequeño y manejable de partes
- Se verifica su correcto funcionamiento una sola vez y puede aplicarse tantas como se necesite
- Código reutilizable por otros programas

La forma de definir una función es:

```
nombre=function(lista_de_argumentos) instrucciones
```

nombre es el nombre de la función, con el cual será llamada en el programa. La *lista_de_argumentos* incluye todas las variables que necesita la función para ejecutarse y que son externas a ella. Las instrucciones procesan la lista de argumentos y generan un resultado que la función devuelve. En el siguiente ejemplo aparece la definición de una función que cálcula el máximo de dos valores y una llamada a la misma

```
maximo=function(a,b){
      if (a>b){return(a)}
      else{return(b)}
}
m=maximo(3,4)
```

En la mayoría de los lenguajes de programación, el valor que devuelve una función se explicita con una orden concreta, generalmente **return**. En S el valor puede devolverse mediante una instrucción **return** pero si no se hace así, la función devuelve la última variable calculada o simplemente ejecutada en el seno de la función. En el anterior ejemplo, si a>b el valor devuelto es a y en caso contrario es b.

A continuación vemos otro ejemplo en el que una función, encargada de resolver ecuaciones de segundo grado, devuelve un vector con las dos soluciones:

```
e2g=function(a,b,c){
   if (a==0){-c/b}
   else{
        r=b\^{}2-4*a*c;
        if (r<0){cat("Las soluciones son complejas");return(c(NA,NA)})
        else{
           s1=(-b+sqrt(r))/(2*a)
           s2=(-b-sqrt(r))/(2*a)
           return(c(s1,s2))
        }
   }
 }
```

Esta función utiliza además dos instrucciones **if ... else** para determinar si el primer parámetro es cero, en cuyo caso se resuelve como una ecuación de primer grado y así se evita que la división por cero posterior de un error, y si las soluciones son complejas. En este último caso se da un mensaje de aviso y se devuelven dos NA.

Para llamarla bastará con ejecutar por ejemplo:

```
> e2g(2,5,3)
```

El resultado será:

```
[1] -1.0 -1.5
```

Con el tiempo, se dispondrá de un elevado número de funciones que pueden agruparse en una librería para ser reutilizadas en diferentes programas. Una librería es un fichero que contiene las definiciones de las diferentes funciones y que se llama al comienzo de un programa con la orden:

```
source("libreria.R")
```

En R existen también librerías de funciones compiladas que se cargan con la función **library**, se pueden descargar como paquetes de la página principal de R.

7.5 Estructura de un programa

Un programa consiste en una secuencia de instrucciones que pueden agruparse en:

- **Entrada de datos**. Conjunto de instrucciones que toman datos de una unidad de entrada, los depositan en la memoria del ordenador. En esta etapa se definen también los **tipos de datos** con los que va a trabajar el programa.
- **Proceso**. Conjunto de instrucciones que resuelven el problema a partir de los datos de entrada y almacenan el resultado en memoria. Describen el algoritmo del programa.

- **Salida**. Conjunto de instrucciones que vuelcan los resultados, a un fichero, pantalla o impresora, con un formato definido por el usuario.

En un programa *profesional* resulta a menudo difícil distinguir unas fases de otras, sin embargo es una buena práctica separar claramente estas etapas cuando se empieza a programar.

Los programas que veremos en los próximos temas implementan modelos sencillos qu normalmente se utilizan enlazados entre si y con otros modelos para crear modelos más complejos.

Cuando se analicen los códigos se separarán claramente las siguientes secciones, aunque no todos los modelos tendrán los mismos elementos.

- **Funciones del modelo**, funciones que ejecutan seperadamente parte del modelo.
- **Parámetros**, magnitudes físicas que van a permanecer constantes en todo el proceso.
- **Variables de ejecución**, no proceden del modelo sino que permiten decidir como se va a ejecutar el programa. Sería por ejemplo la duración de la ejecución en un modelo dinámico.
- **Reserva de memoria para variables de estado**, son las variables que si se modificarán durante la ejecución del programa.
- **Condiciones iniciales**, valores de las variables de estado al inicio del modelo
- **Condiciones de contorno**, valores de las variables de entrada a lo largo de la simulación.
- **Algoritmo**, el meollo de la cuestión, es la parte más compleja pero mno suele ser muy larga.
- **Salidas**, en formato gráfico o de texto.

Ejercicios

1. Haz una función a la que se le pase un número y nos diga si es par o impar
2. Haz un programa en R que imprima los números del 0 al 100, deberás utilizar un bucle `for`
3. Haz un programa que devuelva la suma de los 100 primeros números
4. Realiza una función que imprima la tabla de multiplicar de un número cualquiera que se introducirá en el programa como una variable y escribe un programa que incluya y utilice esta función.
5. Haz un programa que actúe de forma similar a la orden de S `cumsum`, es decir que tome un vector de números X y que genere un vector Y tal que $Y_i = \sum_{j=1}^{i} X_j$
6. Realiza una función a la que se le pasen las longitudes de los tres lados de un triángulo y analice que tipo de triángulo es:

- equilátero
- isósceles
- escaleno
- rectángulo

Recuerda cuales son las *condiciones* que debe cumplir un triángulo de cada uno de estos tipos, si no lo recuerdas consulta algún libro de geometría (muy) básica.

7. Realiza una función que devuelva el número e mediante el desarrollo de la serie:

$$e = 1 + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + ... \tag{7.1}$$

con un error menor que el introducido como dato.

Tema 8

Modelos empíricos

Una de las más clásicas clasificaciones de los sistemas, distingue entre:

- Sistemas abiertos, se producen entradas y salidas de materia y energía
- Sistemas cerrados, se producen entradas y salidas de energía pero no de materia
- Sistemas aislados, no hay transferencias con el entorno, ni de materia ni de energía.

Una de las estrategias de modelización más sencillas consiste en tratar de establecer un modelo matemático que relacione las entradas a un sistema con sus salidas sin pretenden reproducir el comportamiento interno del sistema que se asume irrelevante para los propósitos del modelo. Son los llamados **modelos de caja negra o modelos empíricos**.

Muchos modelos de este tipo se basan en ecuaciones de regresión múltiple en las que la salida que se quiere modelizar es la variable independiente y las entradas y otros factores del sistema son las variables dependientes.

Por ejemplo, podemos suponer que el volumen total mensual de escorrentía de una cuenca depende de la precipitación y de la ETP. Si disponemos de una serie histórica suficientemente larga de las tres variables podremos obtener una ecuación lineal de tipo:

$$r = Ap + Betp + C \tag{8.1}$$

o incluso ecuaciones polinómicas de mayor orden.

Los valores de los parámetros de la regresión (A,B y C) resumen de algún modo un conjunto de factores ambientales propios de la cuenca (tipo de suelo, usos de suelo, pendientes, etc.) y que no se introducen de forma explícita en el modelo.

8.1 Fases en la construcción de un modelo empírico

1. Identificación
2. Calibración
3. Validación
4. Aplicación

8.1.1 Identificación

Basta con representar los datos para tener una idea de que variables pueden actuar como variables independientes en un modelo de regresión y que tipo de ecuación de regresión (lineal, exponencial, etc.) puede ajustarse mejor a los datos (ver figura 6.1).

8.1.2 Calibración

Una vez que se ha identificado el modelo (se dispone de una ecucación, necesitamos obtener los valores de los parámetros que utiliza el modelo. Estos pueden obtenerse utilizando técnicas de **optimización**.

Cuando estamos trabajando con un modelo empírico, los valores de los parámetros deben calibrarse a partir de una muestra de valores de entrada y de salida del modelo y de una función objetivo cuyo valor debe minimizarse.

Uno de los objetivos más sencillos sería un modelo de regresión lineal que utiliza una variable de entrada x y una variable de salida y. El modelo a calibrar sería una ecuación de tipo $\hat{y} = A + B * x$ en el que los parámetros A y B deben ser tales que minimicen la función objetivo $\sum_{i=1}^{n} \hat{y} - y^2$.

Los valores de los parámetros, tras calibrar el modelo, deben tener valores con cierto sentido físico, si no es así puede que el modelo tenga poder predictivo para el conjunto de datos utilizado en la calibración pero tendrá muy poca capacidad explicativa y será muy poco generalizable, por tanto no podrá utilizarse con datos distintos a aquellos que se han utilizado para calibrarlo.

Por ejemplo, en un modelo que relacione temperatura con altitud mediante una ecuación lineal de tipo $t = A + Bz$, el coeficiente B debe ser negativo, sino estaríamos asumiendo que la temperatura aumenta con la altitud lo que es físicamente incorrecto (aunque podría ser cierto en condiciones muy particulares).

8.1.3 Validación y Verificación

Validación es el proceso de comprobar que los resultados aportados por le modelo para las variables de salida y de estado no son muy diferentes a los medidos en la realidad. Existen diferentes índices que permiten cuantificar el grado de ajuste entre los datos observados y los resultados del modelo[1].

[1] En las ecuaciones que siguen, el subíndice o hace referencia a los datos observados, y el subíndice m a los resultados del modelo, $\overline{o}$ es la media de los datos observados

Coeficiente de determinación r^2 , es decir el cuadrado del coeficiente de correlación:

$$r^2 = (\frac{cov(o,m)}{sd(o)sd(m)})^2 \tag{8.2}$$

donde $cov(o,m)$ es la covarianza entre los valores observados y los devueltos por el modelo, $sd(o)$ la desviación típica de los valores observados y $sd(m)$ la desviación típica de los resultados del modelo. Es el más utilizado, oscila entre 0 y 1 y representa el porcentaje de varianza en los datos observados explicado por el modelo.

El problema de este índice es que es insensible a desviaciones constantes o proporcionales, es decir que si se cumple que $m_i = A + B * o_i$, r^2 será igual a 1, aunque $m_i \neq o_i$, haciendonos creer que el modelo responde perfectamente a la realidad. Otro problema es que es muy sensible a los valores extremos que harán crecer el índice dando de nuevo una falsa apariencia de buen ajuste.

Eficiencia del modelo se debe a Nash y Sutcliffe (1970), se basa en la ecuación:

$$NS = 1 - \frac{\sum_{i=1}^{n}(o_i - m_i)^2}{\sum_{i=1}^{n}(o_i - \overline{o})^2} \tag{8.3}$$

Este índice produce resultados menores o iguales a 1, si el resultado es 1 el ajuste es perfecto, si es cero el error es del mismo orden de magnitud que la varianza de los datos observados por lo que la media de los datos observados tendrá una capacidad predictora similar al modelo. Valores inferiores a cero implican que la media tiene una capacidad predictora más alta que el modelo (lo que implica desde luego que el modelo es muy malo).

Este índice no es sensible al efecto de los valores proporcionales pero sigie siendo sensible a los valores extremos.

RMSE/MAE , el cociente entre el error cuadrático medio y el error absoluto medio permite determinar hasta que punto la existencia de valores extremos está afectando al modelo. Cuanto mayor sea el índice mayor será la influencia de los valores extremos.

$$\frac{RMSE}{MAE} = \frac{\sqrt{\frac{\sum_{i=1}^{n}(o_i - m_i)^2}{n}}}{\frac{\sum_{i=1}^{n}|o_i - m_i|}{n}} \tag{8.4}$$

8.1.4 Análisis de sensibilidad

El análisis de sensibilidad mide cuanto pueden llegar a afectar a los resultados de un modelo variaciones relativamente pequeñas en los valores de los parámetros. Tiene un gran número de utilidades:

- En primer lugar sirve para comprobar la lógica interna de un modelo, ayuda a entender como funciona el modelo o porque no funciona correctamente y aprender más acerca de su funcionamiento.

  En un modelo pequeño, con pocos parámetros puede resultar obvio a partir del estudio de sus ecuaciones que parámetros van a tener más influencia sobre los resultados del modelo; pero en un modelo complejo esto no será tan obvio y puede resultar imprescindible un análisis de sensibilidad.

- Para definir la importancia de cada parámetro lo que servirá para deteminar el grado de esfuerzo que debe prestarse a su medición o muestreo.

- Medir como la incertidumbre en la determinación de un parámetros puede afectar a la certidumbre del parámetro.

- Detectar si el modelo está **sobreparametrizado**, esto ocurre cuando existen parámetros a los que el modelo resulta insensible y por tanto no aportan nada a este, en este caso será necesario eliminar algunos para simplificar el modelo.

También puede hacerse un análisis de sensibilidad de las funciones, se trata de determinar como afectan distintas formulación de las ecuaciones utilizadas para modelizar los procesos y relaciones del sistema a los resultados finales.

Uno de los presupuestos básicos del análisis de sensibilidad, y que no tiene por que cumplirse en todos los casos, es que el papel que cada parámetro juega en el modelo es una representación razonable de su papel en el sistema. De este modo la sensibilidad del modelo al parámetro será equivalente a la sensibilidad del sistema al parámetro.

El análisis de sensibilidad suele hacerse ejecutando el modelo para diversos valores del parámetro cuya sensibilidad quiere calcularse dejando fijos todos los demás. Sin embargo la sensibilidad a un parámetro dependerá de los valores adoptados por los demás parámetros, con lo que puede ser más complejo hacer un análisis de sensibilidad. Otra consideración importante es que si el modelo es muy sensible a un parámetro pero este varía muy poco, dicha sensibilidad no va a ser relevante.

En algunos casos puede no ser realista hacer un análisis de sensibilidad ejecutando el modelo para todas las posibles combinaciones de todos los rangos de las variables de interés. En estos casos es preferible utilizar métodos de Montecarlo para hacer unas pocas pruebas con algunos de los posibles valores de las variables a comprobar.

8.2 El modelo de erosión de Thornes(1990)

En primer lugar vamos a ver un modelo sencillo de erosión hídrica del suelo. En este modelo, el conjunto de factores implicados en el proceso se resume en cuatro variables:

- k es el factor K de la USLE (Ecuación Universal de Pérdida de Suelo) y mide la susceptibilidad de un suelo a ser erosionado (erodibilidad), por tanto al aumentar k aumentará la erosión. El cálculo de esta

variable resulta bastante complejo y debe contarse de datos acerca de la textura y el contenido en materia orgánica del suelo;

- Q es el caudal (expresado en mm/mes) que fluye sobre la superficie del suelo. Se trata de una variable que varía tanto en el espacio como en el tiempo, incluye tanto caudal de lluvia como caudal generado aguas arriba. Lógicamente al aumentar el caudal aumenta la erosión;
- S es la pendiente expresada en tantos por 1, al aumentar esta aumenta la erosión
- v es el porcentaje de cubierta vegetal que proteje al suelo de la erosión, por tanto al aumentar la cubierta vegetal disminuye la erosión.

Estas variables se agrupan en el modelo formando la siguiente ecuación:

$$E = kQ^m S^n e^{-iv} \tag{8.5}$$

Los valores de los exponentes deberían ajustarse a cada caso concreto pero una buena aproximación sería: m=1.66, n=2 e i=0.07.

Tanto el coeficiente de erodibilidad como el caudal y la pendiente entran en el modelo multiplicando y con exponente positivo, ya que como se ha visto antes al aumentar estas variables aumentará la erosión. Por el contrario la cobertura vegetal forma parte de un término exponencial negativo, al aumentar esta disminuirá la erosión.

Esta ecuación puede aplicarse a una cuenca o a una celdilla de una capa raster en un SIG. En un programa en R puede introducirse como una función:

```
erosion=function(k,q,s,v){
    E=k*(q^1.66)*(s^2)*exp(-0.07*v)
}

k=2
q=100
s=0.02
v=0.4

E=erosion(k,q,s,v)
```

El proceso de **calibración** parte de valores medidos de las variables dependientes e independientes y busca los valores de los parámetros que, dados los valores de las variables independientes, permitan al modelo obtener una estimación lo más próxima posible a los valores de la variable dependiente. Para ello utilizaremos la función `lm`, pero primero habrá que transformar la ecuación no lineal del modelo de Thornes en una ecuación lineal tomando logaritmos:

$$E = KQ^m S^n e^{-iV} \tag{8.6}$$

$$log(E/K) = mlog(Q) + nlog(S) - iV \tag{8.7}$$

Así que basta hacer las transformaciones de variables:

$$Y = log(E/K) \tag{8.8}$$
$$X1 = log(Q) \tag{8.9}$$
$$X2 = log(S) \tag{8.10}$$
$$X3 = V \tag{8.11}$$

y obtendremos un modelo de la forma:

$$Y = A = B * X1 + C * X2 + D * X3 \tag{8.12}$$

en el que:

$$A = 0 \tag{8.13}$$
$$B = m \tag{8.14}$$
$$C = n \tag{8.15}$$
$$D = -i \tag{8.16}$$

En la función `lm` podemos forzar que A sea igual a 0 escribiendo:

```
m=lm(Y=X1+X2+X3-1)
```

en lugar de

```
m=lm(Y=X1+X2+X3)
```

Podemos hacer un **análisis de sensibilidad** de este modelo tan sólo con definir una de las variables de entrada como un vector, por ejemplo en lugar de un valor de cuadales vamos a utilizar un rango de posibles valores de caudal `q=seq(0,200)`, de esta manera `E` será también un vector de valores de erosión. La representación de ambas variables (en definitiva de la sensibilidad del modelo al caudal, puede hacerse con `plot(q,E,type="l")`.

8.3 Cuestiones

- Calibra el modelo de Thornes con la tabla *thornes.txt* que puedes descargar de la página web de la asignatura.
- ¿Qué conclusión tendría que sacarse si los valores de $Pr(> |t|)$ obtenidos del modelo de regresión fueran superiores al umbral de significación establecido *a priori*?
- Valida el modelo obtenido anteriormente con la tabla de datos *thornes_val.txt* que puedes descargar de la página web de la asignatura.
- Haz un análisis de sensibilidad de los diferentes parámetros y variables
- Si asumimos que la medida de pendiente tiene un error de +-0.02, ¿Cual será el error de la erosión calculada si la pendiente medida es 0.1? ¿y si es 0.9? ¿A que puede deberse esta diferencia?
- Si en lugar de un único valor de caudal utilizaramos una serie temporal, obtendriamos una serie temporal con los valores de erosión correspondientes. A pesar de entrar a considerar el tiempo, en realidad no podría considerarse un modelo dinámico ya que no incluye variables de estado que cambian con el tiempo.

8.4 Modelos para generar series temporales de variables

Las variables de entrada a un modelo no se ven afectadas por el mismo, son por tanto independientes. Por ello los modelos pueden probarse con diferentes tipos de series, datos reales, series inventadas o series obtenidas a partir de modelos de generación de series.

La serie obtenida debe poseer las propiedades estadísticas de la serie real y, en definitiva, esta debería poder confundirse sin problemas con una de las series artificiales.

8.4.1 Series anuales de variables climáticas

Los datos anuales suelen seguir doistribuciones normales por lo que, en principio bastaría con generar una serie de `N` números procedentes de una distribución normal con media P_m y desviación típica P_{sd}:

```
rnorm(N,mean=Pm,sd=Psd)
```

sin embargo las variables ambientales suelen mostrar cierta autocorrelación temporal de manera que el valor de un año se parecerá un poco al del anterior y al del posterior. Este fenómeno es difícil de simular. Una forma sencilla sería generar una primera serie (`S1`)con `rnorm` y hacer que el valor de un año fuese la media ponderada de su valor en S1 y el valor del año anterior (en S1 o en la serie derivada).

El código

```
# Variables de la ejecución

N=100  # Número de años de la serie


# Parámetros del modelo

a=0.4  # Peso de la precipitación en el año anterior


# Reserva de memoria

S1=rnorm(N,mean=300,sd=40)
S2=rep(NA,N);S2[1]=S1[1]
S3=rep(NA,N);S3[1]=S1[1]


# Algoritmo

for (t in 2:N){
  S2[t]=a*S1[t-1]+(1-a)*S1[t]
  S3[t]=a*S2[t-1]+(1-a)*S2[t]
}


# Salidas gráficas

plot(S1,type="l")
lines(S2,col="red")
lines(S3,col="blue")
```

8.4.2 Series mensuales de variables climáticas

Los valores de precipitación mensual proceden de diferentes distribuciones según el mes de que se trate. Por tanto bastaría con un programa similar al siguiente:

```
N=10               # Número de años en la serie
Prec=rep(NA,N*12)  #

t=0
for (ano in 1:N){
    t=t+1;Prec[t]=rnorm(1,mean=20,sd=1)  #Enero
    t=t+1;Prec[t]=rnorm(1,mean=20,sd=1)  #Febrero
    t=t+1;Prec[t]=rnorm(1,mean=20,sd=1)  #Marzo
    t=t+1;Prec[t]=rnorm(1,mean=20,sd=1)  #Abril
    t=t+1;Prec[t]=rnorm(1,mean=20,sd=1)  #Mayo
    t=t+1;Prec[t]=rnorm(1,mean=20,sd=1)  #Junio
    t=t+1;Prec[t]=rnorm(1,mean=20,sd=1)  #Julio
    t=t+1;Prec[t]=rnorm(1,mean=20,sd=1)  #Agosto
    t=t+1;Prec[t]=rnorm(1,mean=20,sd=1)  #Septiembre
    t=t+1;Prec[t]=rnorm(1,mean=20,sd=1)  #Octubre
    t=t+1;Prec[t]=rnorm(1,mean=20,sd=1)  #Noviembre
    t=t+1;Prec[t]=rnorm(1,mean=20,sd=1)  #Diciembre
}
dim(Prec)=c(12,N)
```

El código tal como está planteado no permite fenómenos de autocorrelación que generen episodios de sequía. Sin embargo el procedimiento utilizado en el ejemplo anterior no sirve ahora ya que a escala mensual los períodos secos o húmedos duran varios meses.

Una alternativa sería simular una serie de F(x) (con valores entre 0 y 1) y luego utilizarla con `qnorm` para obtener valores de precipitación adecuados a cada mes.

El siguiente código genera dicha serie como la media ponderada de una onda cuya longitud en meses puede modificarse y un término aleatorio procedente de una distribución uniforme. El peso que se da a la onda es el parámetro `a` mientras que el peso que se da al término aleatorio será `1-a`

Una vez construida la serie de F(x) obtendríamos los correspondientes cuantiles con la función de R `pnorm`.

```
# Variables de ejecución

N=10           # Número de años de la serie
tiempo=N*12    # Número de meses de la serie


#Parámetros

a=0.6          # Peso que se le da a la onda
long=20        # Longitud de onda en meses


# Reserva de memoria

tt=seq(1,tiempo)
Prec=rep(NA,N*12)


# Algoritmo

S1=(sin(6.28*tt/long)+1)/2
S2=runif(tiempo)
S3=a*S1+(1-a)*S2

t=0
for (ano in 1:N){
    t=t+1;Prec[t]=qnorm(S3[t],mean=15,sd=5)  #Enero
    t=t+1;Prec[t]=qnorm(S3[t],mean=12,sd=5)  #Febrero
    t=t+1;Prec[t]=qnorm(S3[t],mean=10,sd=5)  #Marzo
    t=t+1;Prec[t]=qnorm(S3[t],mean=22,sd=5)  #Abril
    t=t+1;Prec[t]=qnorm(S3[t],mean=15,sd=5)  #Mayo
    t=t+1;Prec[t]=qnorm(S3[t],mean=5,sd=5)  #Junio
    t=t+1;Prec[t]=qnorm(S3[t],mean=0,sd=5)  #Julio
    t=t+1;Prec[t]=qnorm(S3[t],mean=2,sd=5)  #Agosto
    t=t+1;Prec[t]=qnorm(S3[t],mean=20,sd=5)  #Septiembre
    t=t+1;Prec[t]=qnorm(S3[t],mean=40,sd=5)  #Octubre
    t=t+1;Prec[t]=qnorm(S3[t],mean=20,sd=5)  #Noviembre
    t=t+1;Prec[t]=qnorm(S3[t],mean=20,sd=5)  #Diciembre
}

# Salidas gráficas y de texto


plot(tt/12,Prec,type="l")
dim(Prec)=c(12,N)
Prec
```

En algunos casos podría ocurrir que Prec[t] adoptara valores negativos. Para corregirlo puede calcularse Prec[t] como:

`max(qnorm(S3[t],mean=20,sd=5),0)`

de manera que los valores inferiores a cero se transforman a 0.

Que ocurre si las condiciones varian?

El modelo anteriormente construido supone por una parte que no hay cambio climático, puesto que las series temporales se generan siempre a partir de la misma función de distribución. Esta hipótesis contradice la existencia de procesos de cambio climático y desertificación.

En el caso de la simulación del cambio climático, se hace necesario permitir modificaciones en los parámetros (media y desviación típica por ejemplo) de las funciones que generan (aleatoriamente) valores para las variables climáticas. Un incremento sostenido de la media producirá un aumento de las variables con el tiempo (y viceversa en el caso de un descenso) mientras que la alteraciones en la desviación típica incrementarán la variabilidad.

Ejercicios

- Reescribe el modelo anterior utilizando los siguientes resultados para Alcantarilla[2]

Mes	Dist.	Parámetros
enero	gamma	scale=38.46 shape=0.639
febrero	gamma	scale=25.64 shape=1
marzo	gamma	scale=27.03 shape=0.982
abril	lognormal	m=2.764 s=1.9
mayo	gamma	scale=25 shape=1.148
junio	gamma	scale=30.3 shape=0.632
julio	gumbel	a=0.148 u=0.357
agosto	gumbel	a=0.143 u=3.113
septiembre	gumbel	a=0.055 u=11.829
octubre	gumbel	a=0.032 u=25.519
noviembre	gamma	scale=30.3 shape=1.011
diciembre	gamma	scale=30.3 shape=0.939

8.4.3 Simulando un episodio de precipitación

El siguiente código simula episodios de precipitación como la yuxtaposición de varios impulsos de lluvia asociados a células tormentosas que se van trasladando sobre el pluviógrafo. Para ello se utilizan las siguientes variables:

- `pep`: Probabilidad de que llegue una de estas células de precipitación
- `mdur` y `sddur`: Media y desviación típica (en minutos) de una distribución normal que produce valores de duración de la célula tormentosa sobre el pluviómetro
- `mprec`: Media de la precipitación caida durante un minuto considerando que haya una célula tormentosa encima (este fenómeno sigue una distribución exponencial negativa

```
#Parámetros

pep=0.2       # Probabilidad de llegada de una célula
mdur=2        # Media de las duraciones (en pulsos) de la célula
```

[2] Alonso Sarría, F. (1995)

```
sddur=1       # Desviación típica de las duraciones (en pulsos) de la célula
mpulso=.5     # Media de la precipitación caida por pulso


# Variables de ejecución

tiempo=60*10  # Duración del episodio en minutos


# Reserva de memoria

prec=rep(0,tiempo)


# Algoritmo

for (t in 1:tiempo){
    if(runif(1)>(1-pep)){
       dur=max(rnorm(1,mean=mdur,sd=sddur),0)
       for (tt in t:(t+dur-1)){
          prec[tt]=prec[tt]+rexp(1,1/mpulso)
        }
     }
}

# Salida gráfica

plot(prec,type="h")
```

Ejercicios

- Trata de entender el código y comentalo
- Genera varios episodios modificando los parámetros y trata de descubrir como afectan estos cambios al resultado

Tema 9

Modelos conceptuales

En ocasiones es posible modelizar sistemas complejos usando ecuaciones lineales simples que, sin ser ecuaciones físicas, representan el funcionamiento del sistema de una forma conceptualmente más cercana a la dinámica del sistema que los modelos empíricos.

9.1 Método racional en hidrología

Es uno de los métodos más utilizados para evaluar la capacidad de producción de caudales de una cuenca fluvial. Fue propuesto por primera vez por Mulvaney en 1850.

Se basa en la ecuación:

$$Q = c * r * A \tag{9.1}$$

donde:

- Q es el caudal, es decir variable de salida;
- r la intensidad de la precipitación, es la variable de entrada y
- c es un coeficiente de escorrentía que mide la proporción de agua de lluvia que no se va a infiltras, es decir que se va a convertir en escorrentía, por tanto es un parámetro
- A el área de la cuenca, otro parámetro.

9.1.1 Agregado y estático

La aplicación directa de esta fórmula en cuencas pequeñas permite obtener el caudal total de la cuenca. Por ejemplo si caen $20 litros/m^2$ en una cuenca de 100 m^2 con un coeficiente de escorrentía de 0.4 el resultado será:

$Q = c * r * A = 0.4 * 20 * 100 = 800$ litros

9.1.2 Semidistribuido y estático

Sin embargo puede utilizarse esta ecuación para conseguir un modelo semidistribuido del comportamiento de una cuenca. Para ello basta con trazar en la cuenca un conjunto de *isocronas*, lineas que unen puntos con el mismo tiempo de concentración[1], de manera que la cuenca quede dividida en sectores en los que el agua caída llegará a la desembocadura a lo largo de un período de tiempo prefijado (digamos una hora).

En la figura 10.1 aparece un ejemplo de división de la cuenca en 5 áreas, donde el área A_i tiene un tiempo de concentración medio de $i - 0.5$ horas, cada una de estas áreas tendrá un coeficiente de escorrentía característico c_i.

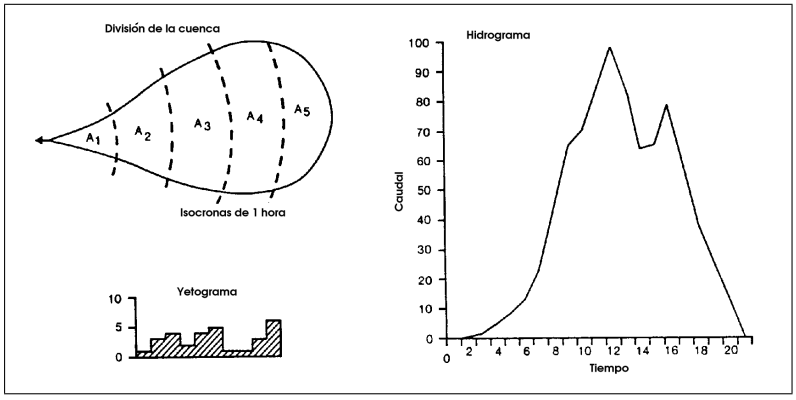


Figura 9.1: Modelo racional

Suponiendo que llueven 10 l/m^2 en una hora, el hidrograma resultante aparece en la tabla 9.1 y la figura 9.2.

Puesto que las cinco áreas en que se divide la cuenca tienen tiempos de concentración medios de 0.5, 1.5, 2.5, 3.5 y 4.5 horas; en la primera hora desaguará la precipitación caída en el área 1 que es igual a $c_1 * r * A_1$, en la segunda hora la precipitación caída en el área 2 y así sucesivamente.

[1] el tiempo de concentración es el tiempo medio que tardará una parte de la cuenca en desaguar por la desembocadura

i	a (m^2)	c (%)
1	2000	40
2	3000	30
3	4000	30
4	3500	20
5	1500	10

t	q (m^3/h)
1	8
2	9
3	12
4	70
5	1.5

Table 9.1: Un ejemplo con el modelo racional

9.1.3 Semidistribuido y dinámico

Podemos convertir el método racional en un modelo dinámico utilizando un yetograma horario en lugar de un único valor de precipitación. Por simplificar asumimos que este yetograma refleja la precipitación en cada hora durante un período de 8 horas, así P_t representa la precipitación caída entre la hora $t-1$ y la hora t que se considera homogenea en toda la cuenca.

Utilizando la ecuación 9.1 podemos concluir que

$$q_{i,t} = c_i * P_t * A_i \tag{9.2}$$

de esta manera calculamos la escorrentía generada en cada una de las i areas en que se ha dividido la cuenca. Si llamamos Q_t al hidrograma de caudales en la desembocadura, podemos escribir:

$$\begin{aligned}
Q_1 &= q_{1,1} \\
Q_2 &= q_{1,2} + q_{2,1} \\
Q_3 &= q_{1,3} + q_{2,2} + q_{3,1} \\
Q_4 &= q_{1,4} + q_{2,3} + q_{3,2} + q_{4,1} \\
Q_5 &= q_{1,5} + q_{2,4} + q_{3,3} + q_{4,2} + q_{5,1} \\
Q_6 &= q_{1,6} + q_{2,5} + q_{3,4} + q_{4,3} + q_{5,2} \\
Q_7 &= q_{1,7} + q_{2,6} + q_{3,5} + q_{4,4} + q_{5,3} \\
Q_8 &= q_{1,8} + q_{2,7} + q_{3,6} + q_{4,5} + q_{5,4} \\
Q_9 &= q_{2,8} + q_{3,7} + q_{4,6} + q_{5,5} \\
Q_{10} &= q_{3,8} + q_{4,7} + q_{5,6} \\
Q_{11} &= q_{4,8} + q_{5,7} \\
Q_{12} &= q_{5,8}
\end{aligned}$$

Abreviando, el caudal de desague en el intervalo t será:

$Q_t = \sum_{i=1}^{t} q_{i,t-i+1}$

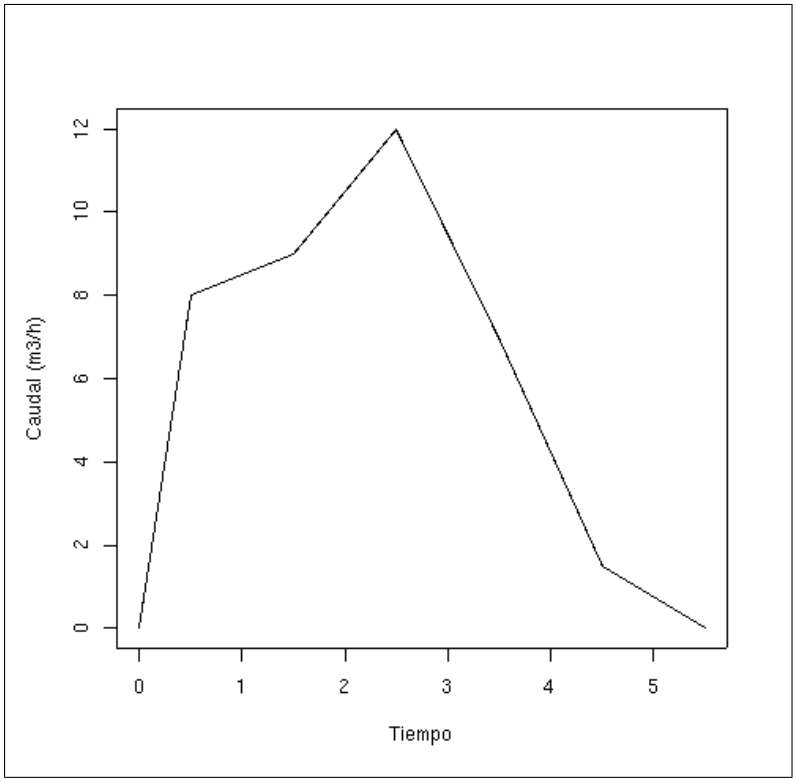


Figura 9.2: Un ejemplo con el modelo racional

9.1.4 El código

```
# Parámetros

    #Areas de las zonas de igual Tc
    a=c(10,15,20,20,25,30,30,20,15,10)

    #Coeficientes de escorrentía
    infi=c(0.40,0.30,0.20,0.30,0.30,0.40,0.50,0.60,0.70,0.80)


# Condiciones de contorno

    #precipitación total
    ptot=100

    #Vector de distribución temporal de la precipitación
    pr=c(0.05,0.075,0.1,0.125,0.15,0.15,0.125,0.1,0.075,0.05)
    r=pr*ptot


# Reserva de memoria de la variable de estado caudal

    q=rep(0,length(a)+length(r)-1)


# Variables de ejecución

    escala_lluvia=50  # Factor por el que se multiplicará la lluvia para generar el yetograma
```

```
#Algoritmo

    for (t in 1:length(r)){
       for (j in 1:length(a)){
        q[t+j-1]=q[t+j-1]+a[j]*infi[j]*r[t]
       }
    }


#Salida gráfica

    plot(q,type="l",xlab="Tiempo",ylab=sprintf("P x %g (mm); Q (m3/h)",escala_lluvia))
    lines(r*escala_lluvia,type="h")
```

9.1.5 Cuestiones

1. Modifica el vector de coeficientes de escorrentía y determina como afecta a los resultados
2. Modifica el vector de distribución temporal de la precipitación y determina como afecta a los resultados
3. ¿A Cual de las dos variables es más sensible el modelo?

Tema 10

Modelos de balance de materia y energía

Los modelos de balace se basan en el principio de conservación de la materia o la energía.

$$I - O = \Delta S \tag{10.1}$$

I es la tasa de entrada de materia (o energía) por unidad de tiempo; O es la tasa de salida de materia (o energía) por unidad de tiempo; y ΔS es el incremento en la cantidad de materia (o energía) almacenado en un determinado componente.

Evidentemente se trata de *cuadrar* los balances, de manera que se tenga una idea clara de en que momentos y en que componentes del sistema está, antes de salir, la materia y energía que ha entrado.

Un buen ejemplo es la famosa ecuación de balance hídrico $P = E + R + I$

10.1 Un modelo de infiltración basado en la ecuación de Green-Ampt

La infiltración del agua de lluvia o escorrentía es un proceso de gran relevancia tanto en hidrología como en ecología, agricultura, etc. Se han propuesto diversos modelos para estudiarla. En general estos modelos asumen que el suelo actúa por una parte absorbiendo agua (como una esponja) y por otro transmitiendola por acción de la gravedad hacia capas inferiores.

La ecuación de Green-Ampt es uno de los modelos de infiltación más utilizados en modelización hidrológica y en hidrología de suelos. Se caracteriza por utilizar entre sus variables independientes el contenido de humedad del suelo por lo que resulta muy apropiada para construir un modelo dinámico. Vamos a utilizar una versión simplificada del modelo de Green-Ampt basado en la ecuación:

$$i_p = K + \frac{S}{Sm} \tag{10.2}$$

la variable independiente es el agua acumulada en el suelo (Sm) y los parámetros son la conductividad hidraúlica saturada K y un parámetro que indica el potencial de absorción de agua del suelo S. Conforme aumenta la humedad del suelo, esta tiende a igualarse a S y el valor de i_p tiende a $K + 1$, es decir conforme disminuye la capacidad de absorción del suelo, este sólo infiltra el agua que puede transmitir a las capas bajas.

Con esta ecuación se obtiene la infiltración potencial (i_p) es decir la máxima capacidad de infiltración del suelo en un instante dado. El utilizar esta ecuación supone la necesidad de integrar un algoritmo que distinga entre infiltración potencial (i_p) e infiltración real (i_r), esta última dependerá no sólo de la capacidad sino también de la cantidad de agua infiltrable (precipitación).

si $P >= ip$ => $ir = ip$ y $e = P - ir$
si $P < ip$ => $ir = P$ y $e = 0$
$Sm = Sm_{t-1} + ir\Delta t$

donde P es la precipitación, e la escorentía , ir la infiltración real y Δt es el intervalo temporal. La obtención de estas variables a partir de la infiltración potencial deberá explicitarse mediante un lenguaje de programación.

```
ip=((S/sm)+K);
if (ip>=p){
    s=s+p;e=0
}
else{
    s=s+ip;e=p-ip
}
```

En la figura 10.1 aparece el diagrama de flujo de este algoritmo.

10.1.1 El código

Para programar este modelo en R se han creado dos funciones (**fsm** y **fr**) que calculan en un instante dado la humedad del suelo (Sm) y la escorrentía (r) a partir de los valores de S, K, Sm (humedad del suelo en el instante anterior) y la precipitación (p).

En segundo lugar se establece la duración de la simulación y se dan valores a los parámetros de succión (S) y conductividad (K)

En tercer lugar se genera una serie de precipitación (p), que actúa en este modelo como variable de entrada, y se reserva espacio en memoria para las variables de estado (Sm) y de salida (r) y se establecen las condiciones iniciales (humedad del suelo en el intervalo inicial).

A continuación se ejecuta el modelo. En un bucle temporal, sucesivas llamadas a las funciones fr y fsm actualizan los valores de la variable de estado (Sm) y de la variable de salida (r).

Finalmente se crea la representación gráfica del modelo.

```
#Funciones del modelo

fsm=function(S,K,sm,p){       #Función humedad del suelo
    i=(S/sm)+K;
    if (i>=p){return (sm+p)}
    else{return (sm+i)}}

fr=function(S,K,sm,p){       #Función escorrentía
    i=(S/sm)+K;
    if (i>=p){return(0)}
    else{return (p-i)}
}


# Variables de ejecución

tott=100      #Tamaño de la simulación

# Parámetros

S=0.1         # Parámetro de succión
K=0.01        # Parámetro de conductividad hidraúlica saturada


#Variables de entrada (Condiciones de contorno)

p=rep(1,tott)

#Reserva de memoria para variables de estado

sm=rep(NA,tott);
r=rep(NA,tott)

# Variables de entrada (Condiciones iniciales)

sm[1]=0.19       #Humedad inicial del suelo


#    ALGORITMO
#    ---------

for (t in 1:(tott-1)){
   cat(t,"\n")
   r[t+1]=fr(S,K,sm[t],p[t+1])
   sm[t+1]=fsm(S,K,sm[t],p[t+1])
}

#    SALIDAS GRAFICAS
#    ----------------
```

```
plot(p,type="h",ylim=c(0,5))
lines(r,col="red")
lines(sm,col="blue")
```

10.1.2 Ejercicios

- Modifica las salidas gráficas anteriores incluyendo leyendas.
- Simula diferentes yetogramas

10.2 Un modelo de balance hídrico

Balance hídrico con el método de Thornwaite[1]

10.2.1 El código

```
# Función para calcular la reserva útil

vru=function(rmax,da){
  rmax*exp(da/rmax)
}

# Variables de entrada (Condiciones de contorno)

p=c(70,71,60,41,26,7,1,3,24,45,75,74)           # Precipitación
etp=c(14,18,35,52,83,123,164,148,101,60,28,14)  # ETP
rmax=100                                        # Reserva máxima

# Variables de entrada (Condiciones iniciales)

presumd=0                                       # Valor inicial del sumatorio del déficit
preru=100                                       # Valor inicial de la reserva útil

# Reserva de memoria para variables de estado y de salida

sumd=rep(0,12)   # Sumatorio del déficit
ru=rep(0,12)     # Reserva útil
etr=rep(0,12)    # ETR
esco=rep(0,12)   # Escorrentía
sup=rep(0,12)    # Superavit (P-ETP)

# Algoritmo
# ---------

for (m in 1:12){
    if (m==1){rupre=preru;sumdpre=presumd}else{rupre=ru[m-1];sumdpre=sumd[m-1]}
    sup[m]=p[m]-etp[m]
    if (sup[m]<0){
         sumd[m]=sumdpre+sup[m]
         ru[m]=vru(rmax,sumd[m])
         etr[m]=p[m]+(ru[m-1]-ru[m])
         esco[m]=0
    }
    else{
         etr[m]=etp[m]
         ru[m]=min(ru[m-1]+sup[m],rmax)
         if(ru[m]<rmax){esco[m]=0}else{esco[m]=sup[m]-(ru[m]-rupre)}
    }
}
```

[1] Fernandez García, F. (2005) Manual de Climatología Aplicada (cap. 9)

```
# Salidas
# -------

res=data.frame(p,etp,etr,ru,esco)

plot(etp,type="l")
lines(etr,col="red")
lines(p,col="blue")
lines(ru,col="green")
legend(10,150,legend=c("ETP","ETR","P","RU","Esc"),fill=c("black","red","blue","green","cyan"))
```

Este modelo clásico de balance hídrico muestra la necesidad de definir condiciones iniciales y de contorno en un modelo dinámico. Muchas veces resulta problemático establecer unas condiciones iniciales, esto es particularmente cierto con este modelo (normalmente el algoritmo se empieza a ejecutar en el primer mes posterior al período seco asumiendo que la reserva es cero). Una alternativa más eficaz sería ejecutar el modelo varias veces utilizando como condiciones iniciales las condiciones finales de la ejecución anterior. De esta manera al cabo de varias ejecuciones las condiciones iniciales habrían dejado de tener peso.

El algoritmo para hacerlo así quedaría:

```
Nejec=5
for (a in 1:Nejec){
   for (m in 1:12){
       if (m==1){rupre=preru;sumdpre=presumd}else{rupre=ru[m-1];sumdpre=sumd[m-1]}
       sup[m]=p[m]-etp[m]
       if (sup[m]<0){
            sumd[m]=sumdpre+sup[m]
            ru[m]=vru(rmax,sumd[m])
            etr[m]=p[m]+vru(rmax,sumd[m])
            esco[m]=0
       }
       else{
            etr[m]=etp[m]
            ru[m]=min(ru[m-1]+sup[m],rmax)
            if(ru[m]<rmax){esco[m]=0}else{esco[m]=sup[m]-(ru[m]-rupre)}
       }
   }
   presumd=sumd[12];preru=ru[12]
}
```

Simplemente se ha introducido el anterior algoritmo en un bucle en el que se define el número de ejecuciones (`Nejec`) y al final se cada ejecución se actualizan las condiciones iniciales a los valores de las condiciones finales de la ejecución anterior.

El modelo clásico de balance hídrico crea una serie infinita en la que a cada mes se asignan los valores medios de ese mes ($P = P_m$). Este enfoque es muy poco realista, especialmente en climas áridos) y además no permite simular como se comporta el sistema en caso de episodios extremos de precipitación o sequía.

10.2.2 Ejercicios

- Añade una función para calcular la ETP por el método de Thornwaite[2].
- Utiliza el modelo con el conjunto de datos que tu quieras y genera salidas gráficas adecuadas.
- ¿Cómo modificarías el algoritmo para utilizar, en lugar de valores medios, series temporales?

[2]Fernandez García, F. (2005) Manual de Climatología Aplicada (cap. 8)

- Este modelo podría utilizarse para simular el proceso de desertificación. Para ello bastaría con hacer que el parámetro de reserva máxima se convirtiera en variable. ¿Cómo lo harías?

10.3 El mundo de las margaritas

El mundo de las margaritas (*Daisyworld*) es un modelo sencillo para el estudo del efecto de retroalimentación entre planta y clima (Watson and Lovelock 1983). Fue desarrollado como respuesta a las criticas recibidas por la Hipótesis Gaia de J. Lovelock en la que se mantenía que la Tierra actúa como una entidad con capacidad autorreguladora por parte de los seres vivos para mantener unas condiciones adecuadas para la vida (Lovelock 1995a)

El modelo asume un planeta habitado sólo por dos tipos de margaritas (blancas y negras), las primeras con un albedo mayor que el del suelo y las segundas con un albedo inferior. Las margaritas pueden desarrollarse con temperaturas entre 5^OC y 40^OC con un óptimo de crecimiento en 22.5^OC. Los valores de la temperatura dependen del sol y del albedo (a su vez influenciado por las margaritas.

10.3.1 El código

```
# FUNCIONES
# ---------

#Devuelve el área ocupada en t a partir de la ocupada en t-1
#     area = area ocupada en t-1
#     areasuelo = area desnuda en t-1
#     br = tasa de nacimientos
#     dr = tasa de defunciones

nuevo=function(area,areasuelo,br,dr){
  dareadt=(area*(areasuelo*br-dr))
  area+dareadt
}

# VARIABLES DE EJECUCION
# ----------------------

tiempo_ejec=500           #Duración de la ejecución
tiempo=seq(1,tiempo_ejec)

# VARIABLES DE CONTORNO: CONSTANTE SOLAR
# --------------------------------------

S=rep(0,tiempo_ejec)
S[tiempo]=1
#S[tiempo]=0.8+sin(tiempo/100)/5
#S[tiempo]=0.8+tiempo/2500



# PARAMETROS
# ----------
albedo_soil=0.5             #Albedo del suelo desnudo
albedo_black=0.25           #Albedo de las margaritas negras
albedo_white=0.75           #Albedo de las margaritas negras
sigma=5.67*10^(-8)          #Constante de Stefan
L=1000                      #Luminosidad del sol
dr=0.2                      #Tasa de mortalidad de las margaritas


#Declaración de las variables de estado
area_black=rep(NA,tiempo_ejec)       #Porcentaje del planeta con margaritas negras
area_white=rep(NA,tiempo_ejec)       #Porcentaje del planeta con margaritas blancas
area_soil=rep(NA,tiempo_ejec)        #Porcentaje del planeta sin margaritas
```

```
Temp=rep(NA,tiempo_ejec)             #Temperatura media planetaria
Temp_nd=rep(NA,tiempo_ejec)          #Temperatura sin margaritas
Tb=rep(NA,tiempo_ejec)               #Temperatura con margaritas negras
Tw=rep(NA,tiempo_ejec)               #Temperatura con margaritas blancas
brw=rep(NA,tiempo_ejec)              #Tasa de natalidad de las margaritas blancas
brb=rep(NA,tiempo_ejec)              #Tasa de natalidad de las margaritas negras
albedo_global=rep(NA,tiempo_ejec)    #Albedo global


# CONDICIONES INICIALES
# ---------------------

area_black[1]=0.2
area_white[1]=0.2


# ALGORITMO
# ---------

for (t in tiempo){
   area_soil[t]=1-(area_black[t]+area_white[t])
   albedo_global[t]=(area_soil[t]*albedo_soil)+(area_black[t]*albedo_black)+
                    (area_white[t]*albedo_white);
   Temp[t] =(S[t]*L*(1-albedo_global[t])/sigma)^(0.25)- 273.2
   Temp_nd[t]=(S[t]*L*(1-albedo_soil)/sigma)^(0.25)- 273.2

   Tb[t] = (20*(albedo_global[t]-albedo_black) + Temp[t])
   Tw[t] = (20*(albedo_global[t]-albedo_white) + Temp[t])

   brb[t] =(1-(0.003265*(22.5-Tb[t])^2))
   brw[t] =(1-(0.003265*(22.5-Tw[t])^2))

   area_black[t+1]=nuevo(area_black[t],area_soil[t],brb[t],dr)
   area_white[t+1]=nuevo(area_white[t],area_soil[t],brw[t],dr)
}


# SALIDAS GRAFICAS
# ----------------

plot(area_black,type="l",
     xlab="Tiempo",
     ylab="Cobertura y T/100",
     ylim=c(0,1.2))

lines(area_white, col="yellow")
lines(area_soil, col="brown")
lines(S,col="green")
lines(Temp/100,col="blue")
lines(Temp_nd/100,col="red")

legend(tiempo_ejec/2,
       1.2,
       legend=c("Constante Solar",
                "Temperatura sin margaritas",
                "Temperatura con margaritas",
                "Margaritas blancas",
                "Margaritas negras",
                "Suelo desnudo"),
       fill=c("green",
              "red",
              "blue",
              "yellow",
              "black",
              "brown"))
```

10.3.2 Cuestiones

1. El código incluye tres escenarios alternativos por lo que se refiere a la evolución de la constante solar (fija, creciente y oscilante). Comprueba lo que ocurre con los diferentes escenarios.

2. Intenta crear otros escenarios nuevos y comprueba sus efectos

3. Intenta dar una explicación ecólogica a los diferentes comportamientos.
4. Por encima y debajo de que valores de constante solar se produce la extinción

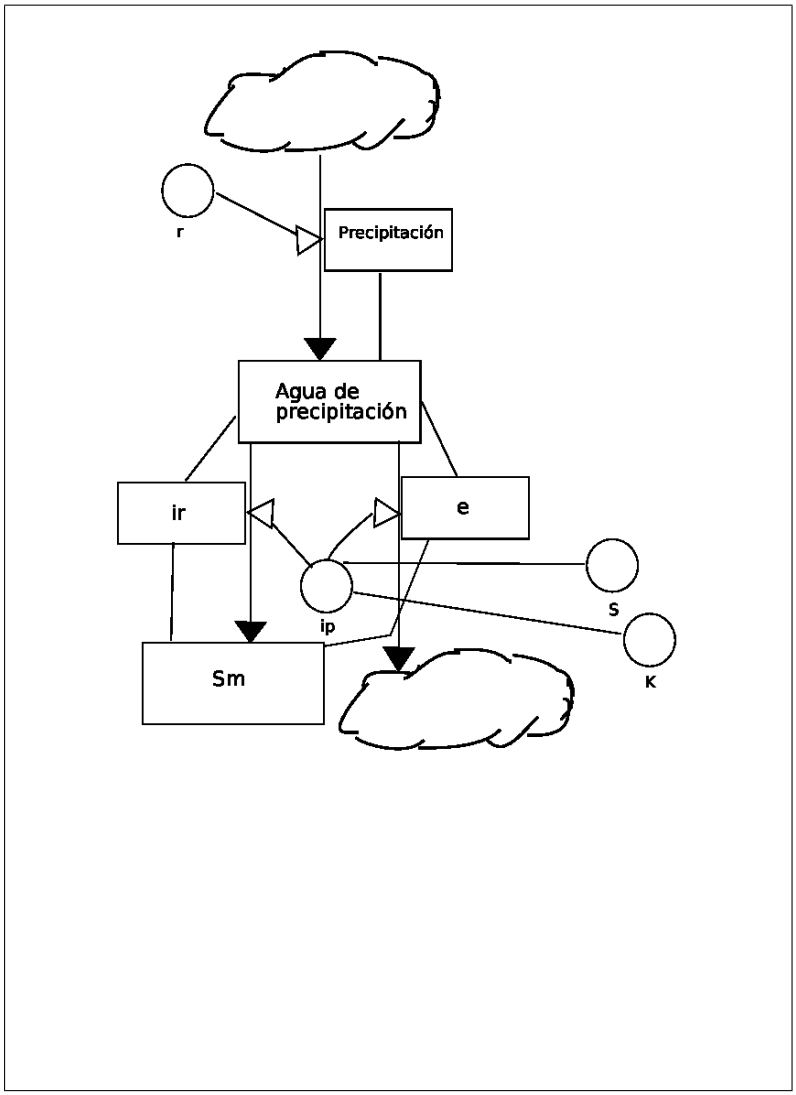


Figura 10.1: Modelo de infiltración a partir de la ecuación de Green-Ampt

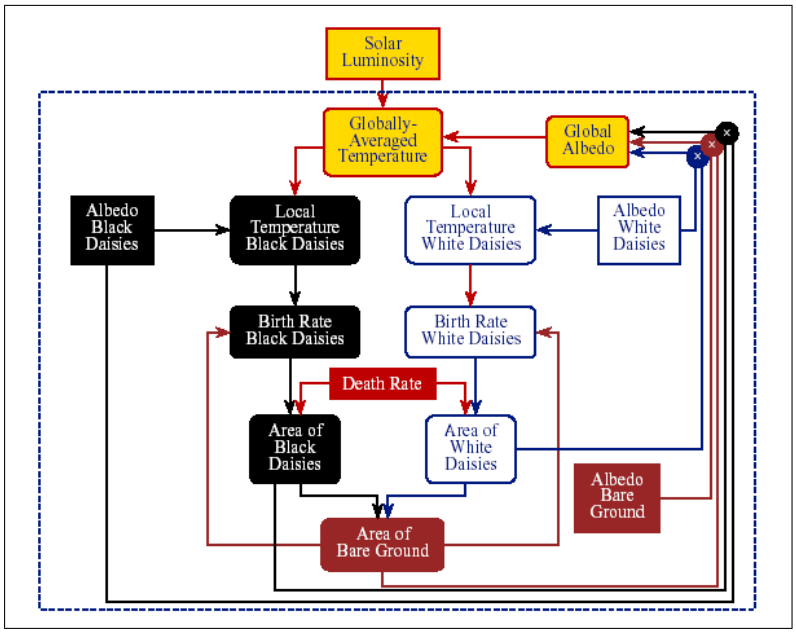


Figura 10.2: Modelo del mundo de las margaritas

Tema 11

Modelos de base física

Los modelos de base física se basan en la representación del ristema mediante ecuaciones matemáticas que representan las leyes que los gobiernan. Normalmente se trata de **ecuaciones diferenciales** que representan la variación de una variable respecto a otra cuando esta última se hace infinitamente pequeña. Por ejemplo la velocidad es la variación del espacio recorrido respecto al tiempo:

$$v = \frac{\delta E}{\delta t} \tag{11.1}$$

En los casos más simples, estas ecuaciones diferenciales pueden resolverse mediante técnicas de cálculo (muy complejas para los no matemáticos); pero los casos más complejos (como puede ser la modelización de procesos ambientales) no pueden resolverse mediante dichas técnicas y deben utilizarse **métodos numéricos**, algo menos exactos, que, irónicamente, resultan mucho más sencillos para un no matemático. Se trata de sustituir las variaciones infinitesimales por incrementos finitos:

$$v = \frac{\Delta E}{\Delta t} = \frac{E_t - E_{t-1}}{\Delta t} \tag{11.2}$$

luego podemos obtener el punto en el que nos encontramos a partir del punto en que nos encontrábamos en el instante anterior.

$$E_t = E_{t-1} + v\Delta t \tag{11.3}$$

Por otro lado, resultaría imposible muestrear y modelizar los sistemas ambientales con precisión infinita, por el contrario lo que se hace es discretizar tanto el espacio como el tiempo. Un buen ejemplo son los SIG raster que establecen una discretización del espacio (tamaño de celdilla) que se aplica a todas las variables representadas.

Finalmente hay que recordar que suele definirse a un ordenador como una *máquina de estados finitos* es decir que no puede procesar fenómenos continuos a no ser que se discreticen adecuadamente.

11.1 Modelos hidrológicos distribuidos de tipo físico: Un modelo de onda cinemática

El flujo de agua y materiales en el espacio y a través del tiempo, es gobernado por una serie de principios básicos.

- Conservación de la masa (ecuación de la continuidad):

$$\frac{\delta Q}{\delta x} + \frac{\delta A}{\delta t} = q \tag{11.4}$$

  donde q son las entradas al sistema desde el exterior por unidad de longitud.

- Conservación del momento (segunda ley de Newton) que en su forma simplificada puede escribirse:

$$A = \alpha Q^{\beta} \tag{11.5}$$

- Conservación de la energía: Primera ley de la termodinámica

$$\Delta E = I_c - W \tag{11.6}$$

Dejando al margen la conservación de la energía, el proceso queda simplemente regido por dos ecuaciones que se pueden integrar en un único modelo. En estas ecuaciones, Q es el caudal, A es el área de la sección mojada, δt el intervalo de tiempo, δx el intervalo espacial equivalente a la resolución o tamaño de la celdilla. Finalmente, α y β son dos parámetros que dependen del tipo de formulación que se emplea para relacionar caudal con área de la sección mojada. Si se emplea la formulación de Manning:

$$\alpha = (\frac{nP^{2/3}}{\sqrt{S}})^{0.6} \tag{11.7}$$

$$\beta = 0.6 \tag{11.8}$$

Diferenciando la ecuación 11.5 y sustituyendo $\frac{\delta A}{\delta t}$ por su equivalente según la ecuación 11.4 se obtiene:

$$\frac{\delta Q}{\delta x} + \alpha\beta Q^{\beta-1}\frac{\delta Q}{\delta t} = q \tag{11.9}$$

Esta ecuación corresponde al modelo de onda cinemática, una simplificación de las ecuaciones de SaintVenant (Chow *et al.*, 1984) y puede resolverse mediante un procedimiento de diferencias finitas haciendo las siguientes sustituciones:

$$\frac{\delta Q}{\delta x} = \frac{Q_{j+1,i+1} - Q_{j+1,i}}{\Delta x} \tag{11.10}$$

$$\frac{\delta Q}{\delta t} = \frac{Q_{j+1,i+1} - Q_{j,i+1}}{\Delta t} \tag{11.11}$$

$$Q = \frac{Q_{j+1,i} + Q_{j,i+1}}{2} \tag{11.12}$$

$$q = \frac{q_{j+1,i+1} + Q_{j,i+1}}{2} \tag{11.13}$$

Los subíndices j e i hacen referencia a los diferentes intervalos temporales y espaciales, respectivamente, según el esquema de la figura 11.2. Esta representa una malla espacio-temporal que simboliza el proceso de resolución del modelo de modo que el cauda en el punto $j+1, i+1$ se resuelve a partir de los valores ya conocidos en $j, i+1$, $j+1, i$ y j, i. Esta resolución requiere conocer:

- Las condiciones iniciales, es decir el valor de caudal en todos los puntos i cuando $j = 0$.
- Las condiciones de contorno o hidrograma de entrada al sistema, es decir el valor del caudal en $i = 0$ para todo j
- Las entradas al sistema (precipitación efectiva), es decir los valores de q para todo i y j

Sustituyendo y despejando para $Q_{j+1,i+1}$ se obtiene:

$$Q_{j+1,i+1} = \frac{(\frac{\Delta t}{\Delta x} Q_{j+1,i} + \alpha\beta Q_{j,i+1}(\frac{Q_{j,i+1}+Q_{j+1,i}}{2})^{\beta-1} + \Delta t(\frac{q_{j+1,i+1}+q_{j,i+1}}{2}))}{\frac{\Delta t}{\Delta x} + \alpha\beta(\frac{Q_{j,i+1}+Q_{j+1,i}}{2})^{\beta-1}} \tag{11.14}$$

Este algoritmo asume un cauce lineal que se divide en intervalos discretos, en el caso de una cuenca tenemos una cuenca bidimensional, pero puesto que cada celdilla recibe el flujo de sus celdillas tributarias (figura 11.3), el valor de $Q_{j,i}$ y $Q_{j+1,i}$ se calcula como la suma de los caudales de las celdillas tributarias

El inverso del parámetro $\frac{\Delta t}{\Delta x}$ equivale a la máxima velocidad que puede alcanzar el flujo para que la resolución del modelo conserve la estabilidad. Se trata de la condición de Courant que establece que:

$$V_{max} < \frac{\Delta x}{\Delta t} \tag{11.15}$$

y por tanto

$$\Delta x > V_{max} \Delta t \tag{11.16}$$

ya que en otro caso el sistema se vuelve inestable. Una forma intuitiva de entenderlo es que si no se cumpliese se daría la paradoja de que $V_{max}\Delta t$, que es el espacio recorrido por un determinado volumen de agua entre dos intervalxos de tiempo del modelo, sería mayor que δx y por lo tanto el agua habría saltado de una celdilla a otra sin pasar por la intermedia.

La condición de Couran implica que las escalas espaciales y temporales de los modelos físicos estén uy vinculadas. Si se quiere trabajar con un modelo de elevaciones detallado, debe hacerse con intervalos temporales bajos.

11.1.1 El código

```
# FUNCIONES DEL MODELO
# --------------------

# q2y_rect. Devuelve la altura de la lámina de agua en un cauce rectangular
#           q=caudal (m3/s)
#           b=anchura (m)
#           s=pendiente (tantos por uno)
#           n=número de Manning

q2y_rect=function(q,b,s,n){
   return((n*q/(1.49*sqrt(s)*b))^(3/5))
}

# alfa. Calcula el parámetro alfa del modelo de onda cinemática
#           s=pendiente (tantos por uno)
#           p=perímetro mojado (m)
#           n=número de Manning

alfa=function(n,p,s){
   return((n*p^(2/3)/(1.49*sqrt(s)))^0.6)
}


# Variables de ejecución

tiempo=100;tramos=10

dx=3000;dt=180

# Parámetros del cauce

s=c(0.09,0.08,0.07,0.06,0.05,0.04,0.03,0.02,0.01,0.005)
n=rep(0.035,10)
w=c(20,30,40,50,60,70,80,90,100,120)

# Reserva de memoria para las variables espacio-temporales

qq=rep(0,tiempo*tramos);dim(qq)=c(tiempo,tramos)
q=rep(0,tiempo*tramos);dim(q)=c(tiempo,tramos)
y=rep(0,tiempo*tramos);dim(y)=c(tiempo,tramos)
p=rep(0,tiempo*tramos);dim(p)=c(tiempo,tramos)


# Definición del hidrograma de avenida (condición de contorno)

qq[,1]=2000
qq[1,1]=2000;qq[6,1]=6000;
qq[2,1]=2000;qq[7,1]=5000;
qq[3,1]=3000;qq[8,1]=4000;
qq[4,1]=4000;qq[9,1]=3000;
qq[5,1]=5000;qq[10,1]=2000;
qq[11,1]=2000
t=1:500

# Definición de las aportaciones exteriores (condición de contorno)

# Valores espaciotemporales en el primer tramo

for (i in t){
```

```
   y[i,1]=q2y_rect(qq[i,1],w[1],s[1],n[1])
   p[i,1]=w[1]*y[i,1]
}

# Definición del caudal base (condición inicial)

qq[1,]=2000

# Variables espaciotemporales en el primer instante

x=1:tramos
for (i in x){
   y[1,i]=q2y_rect(qq[1,i],w[1],s[1],n[1])
   p[1,i]=w[1]*y[1,i]
}


# Algoritmo

for (cx in 2:tramos){
    for (ct in 2:tiempo){
        a=alfa(n[cx],p[ct-1,cx],s[cx])
        b=0.6
        t1=qq[ct,cx-1]*dt/dx
        t2=a*b*qq[ct-1,cx]
        t3=((qq[ct-1,cx]+qq[ct,cx-1])/2)^(b-1)
        t4=dt*(q[ct,cx]+q[ct-1,cx])/2
        t5=(dt/dx)+a*b*t3
        qq[ct,cx]=(t1+t2*(t3+t4))/t5
        y[ct,cx]=q2y_rect(qq[ct,cx],w[cx],s[cx],n[cx])
        p[ct,cx]=w[cx]*y[ct,cx]

        #cat(n[cx]," ",p[ct-1,cx]," ",s[cx]," ",w[cx],"\n")
        #cat(cx,ct," ",a,b," ",qq[ct,cx-1],qq[ct-1,cx],"  ",t1,t2,t3,t5,"  ",qq[ct,cx],"\n")
        #cat(qq[ct,cx], y[ct,cx],p[ct,cx],"\n")
      }
}


# Salida
# ------
t=(1:tiempo)*dt/3660
plot(t,qq[,1],type="l",xlab="Tiempo (h)",ylab="Q (m3/s)")
lines(t,qq[,5],col="red")
lines(t,qq[,10],col="blue")
legend(4,4000,legend=c("X=1","X=5","X=10"),fill=c("black","red","blue"))
```

11.1.2 Ejercicios

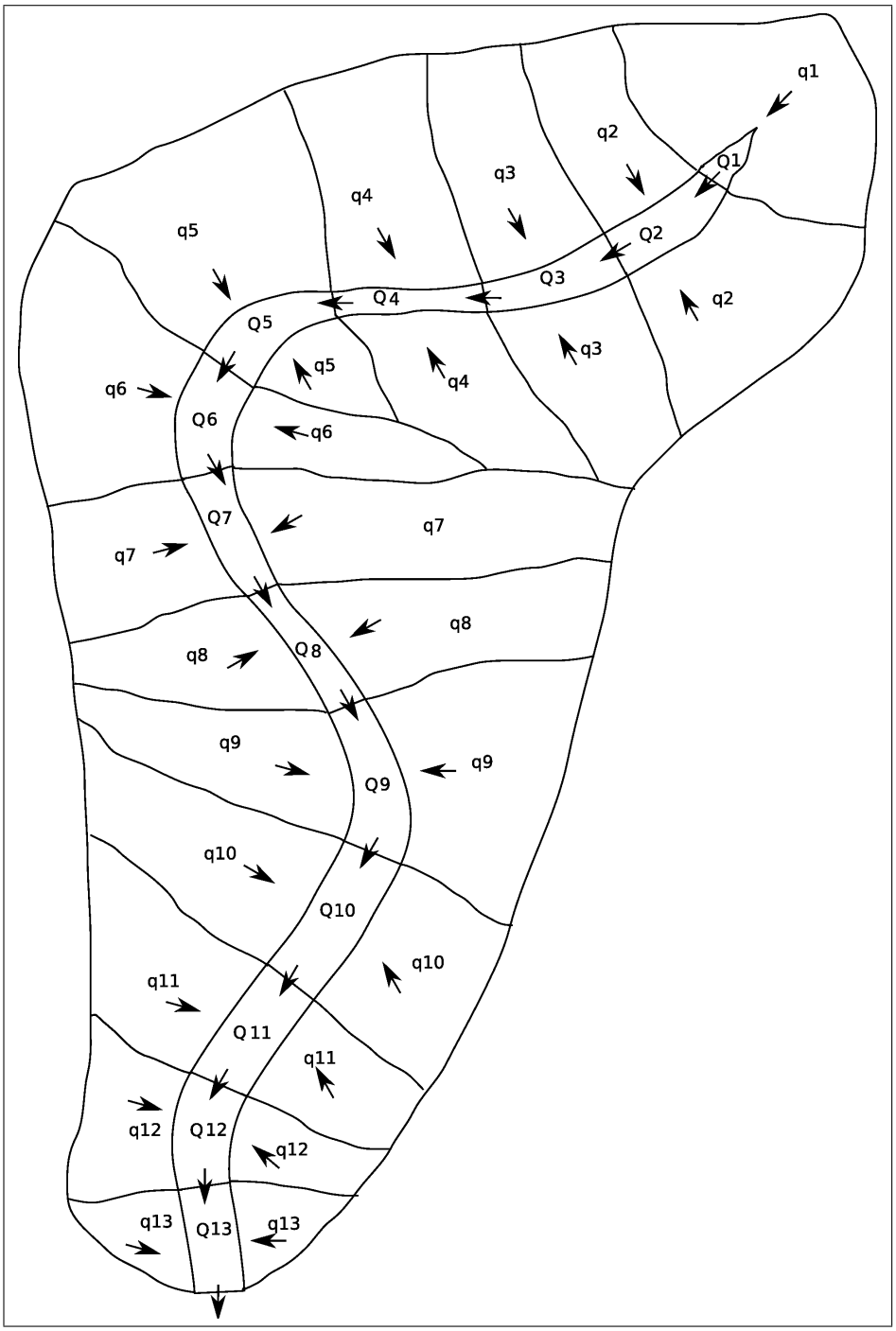


Figura 11.1: Esquema de la relación de los sectores de una cuenca con el cauce principal

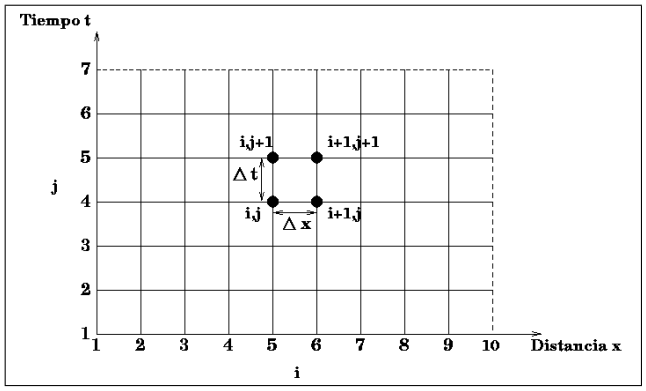


Figura 11.2: Dominio espacio-tiempo para resolver el flujo en un cauce. El caudal y altura de agua se calculan en cada punto de la malla (i+1,j+1) a partir de los valores en los puntos anteriores tanto en el espacio como en el tiempo (i,j; i+1,j; i,j+1).Basdo en Chow *et al.*, (1994)

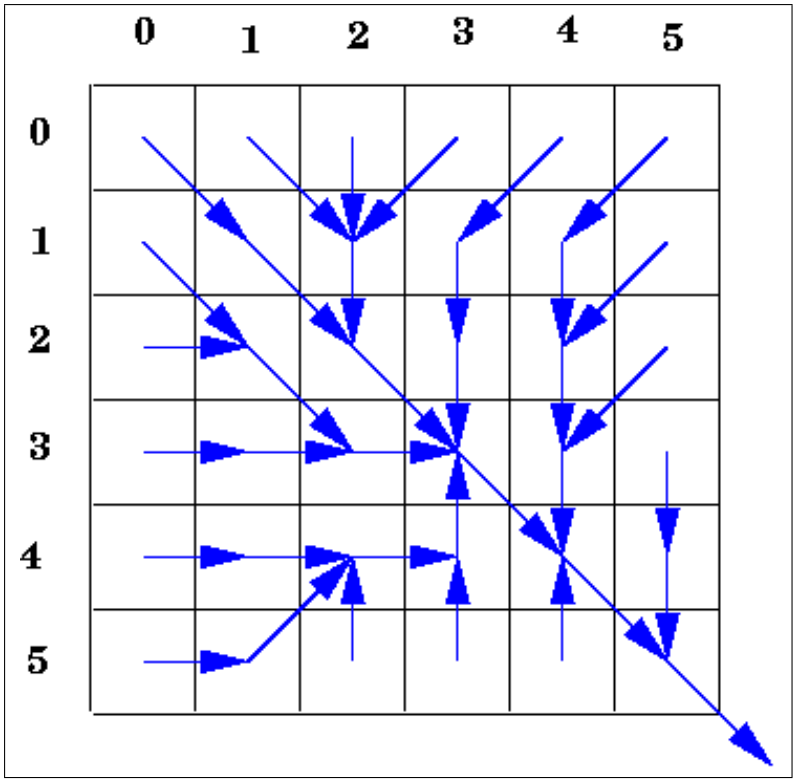


Figura 11.3: Dominio espacial en 2 dimensiones