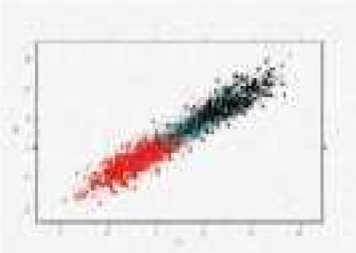
ESTADÍSTICA I (SISTEMAS)

Profesores: Hilario Navarro. Jorge Martín



DEPARTAMENTO DE ESTADÍSTICA,
INVESTIGACIÓN OPERATIVA Y CÁLCULO
NUMÉRICO

Soluciones a los problemas del examen.
Primera prueba presencial. Curso 2002-2003

Problema 1. Primera semana. Extraemos tres dígitos al azar entre el 0 y el 9 para formar una clave. Se pide:

(a) Calcular la probabilidad de que la clave tenga al menos dos cifras iguales.

(b) Calcular la probabilidad de que, si la clave obtenida es un número par, no sea superior a 100.

Solución

(a) Se pueden formar un total de 10^3 claves con los dígitos del 0 al 9.

Denotaremos por A el suceso

$$A = \text{“la clave tiene al menos dos cifras iguales”}$$

Vamos a calcular la probabilidad del complementario

$$A^c = \text{“todas las cifras que forman la clave son distintas”}$$

Hay $10 \cdot 9 \cdot 8$ claves favorables al suceso A^c; ya que la cifra de las centenas puede ser uno cualquiera de los diez dígitos, la de las decenas uno de los nueve restantes y la de las unidades uno cualquiera de los ocho que no ocuparon el lugar de las centenas y decenas.

$$\overset{10}{\boxed{\cdot}}\quad \overset{9}{\boxed{\cdot}}\quad \overset{8}{\boxed{\cdot}}$$

La probabilidad del complementario es $P(A^c) = \dfrac{10 \cdot 9 \cdot 8}{10^3}$; de donde se sigue la probabilidad pedida:

$$P(A) = 1 - P(A^c) = 1 - \frac{10 \cdot 9 \cdot 8}{10^3} = 1 - \frac{18}{25} = \frac{7}{25}.$$

(b) Vamos a resolver el problema utilizando dos métodos.

Método 1. El enunciado nos informa sobre el resultado del experimento: la clave obtenida es un número par. Con esta información la incertidumbre se modifica; de entrada excluiríamos todas las claves impares. Por tanto, el espacio muestral cambia y queda restringido al conjunto de todas las claves pares entre la $\boxed{0}\boxed{0}\boxed{0}$ y la $\boxed{9}\boxed{9}\boxed{9}$; un total de 500, es decir

$$\Omega = \{\text{Conjunto de claves pares entre la } \boxed{0}\boxed{0}\boxed{0} \text{ y la } \boxed{9}\boxed{9}\boxed{9}\}$$

De todas ellas hay un total de 51 que no superan a 100; todos los pares comprendidos entre el 0 cuya clave es $\boxed{0}\boxed{0}\boxed{0}$ y el 100 con clave $\boxed{1}\boxed{0}\boxed{0}$. Consecuentemente, la probabilidad pedida será 51/500.

Método 2. Se considera el espacio muestral inicial que está formado por el conjunto de todas las claves comprendidas entre el 0 y el 999:

$$\Omega = \{\text{Conjunto de claves entre la } \boxed{0}\boxed{0}\boxed{0} \text{ y la } \boxed{9}\boxed{9}\boxed{9}\}$$

A continuación, se consideran los sucesos

$$A = \text{“la clave obtenida no supera a 100”}$$

$$B = \text{“la clave obtenida es un número par”}$$

Nos están pidiendo calcular la probabilidad condicionada

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

Del conjunto de todas las claves, un total de 1000, hay 51 que son pares menores o iguales que 100; con lo cual se tiene que $P(A \cap B) = \dfrac{51}{1000}$.
Por otro lado, hay un total de 500 claves que son pares; luego se obtendrá que $P(B) = \dfrac{500}{1000}$.
Por tanto, la probabilidad pedida es

$$P(A|B) = \frac{51/1000}{500/1000} = \frac{51}{500}.$$

□

Problema 2. Primera semana. La lectura registrada por cierto aparato de medida es una variable aleatoria X con distribución normal. En una muestra de doce lecturas, tomadas al azar, se han observado los siguientes valores de la variable X.

10 15 11 12 8 13 16 5 14 5 6 5

(a) Dar una estimación puntual de la media de X. Justifíquese la elección del estimador que se utiliza.

(b) Obtener un intervalo de confianza, al 95 %, para la media de X.

Solución

(a) Para estimar la media de X, utilizamos la media muestral: $\bar{X} = \dfrac{\sum_{i=1}^{n} X_i}{n}$.

Se propone este estimador porque es insesgado para el parámetro media de la variable poblacional X que se desea estimar. Con la muestra obtenida, la estimación resulta:

$$\bar{x} = \frac{10+15+11+12+8+13+16+5+14+5+6+5}{12} = 10$$

(b) Puesto que la variable aleatoria poblacional es normal de varianza desconocida, la cantidad pivotal que utilizamos para construir el intervalo de confianza pedido es la basada en la distribución t de Student, que elimina el efecto de la varianza poblacional:

$$U = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

donde

- μ es la media de la variable X.
- $\bar{X}$ es la media muestral. En este caso la media muestral es 10.
- S^2 es la cuasivarianza muestral. En la muestra obtenida $S = 4.1$.
- n es el tamaño muestral. En este caso $n = 12$.

En primer lugar, encontramos el intervalo que contiene a U con una probabilidad igual al nivel de confianza 0.95. Puesto que U tiene distribución t con 11 grados de libertad, el problema se reduce a buscar los extremos de un intervalo que contenga a una distribución t_{11} con una probabilidad de 0.95. Los extremos del intervalo, obtenidos de las tablas de la distribución t_{11}, son $-t_{11;0.025} = -2.201$ y $t_{11;0.025} = 2.201$; ya que, por la simetría de la densidad de la distribución t (véase la figura 1), la probabilidad del intervalo $(2.201, \infty)$ es 0.025, la del $(-\infty, -2.201)$ es 0.025 y la del intervalo $(-2.201, 2.201)$ es 0.95 (área sombreada de la figura 1).

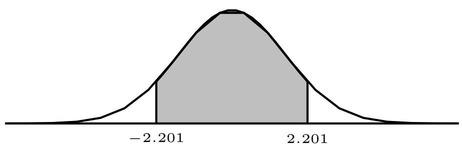


Figura 1: Función de densidad de una t_{11}

Ahora que conocemos el intervalo, el resto del ejercicio es un sencillo cálculo matemático. Dado que

$$0.95 = P(-2.201 \leq \frac{\bar{X}-\mu}{S/\sqrt{12}} \leq 2.201) = P(-2.201\frac{S}{\sqrt{12}} \leq \bar{X}-\mu \leq 2.201\frac{S}{\sqrt{12}}),$$

despejando en la desigualdad anterior la media μ, se obtiene que

$$\left(\bar{X} - 2.201\frac{S}{\sqrt{12}}\ ,\ \bar{X} + 2.201\frac{S}{\sqrt{12}}\right)$$

es un intervalo aleatorio que cubre a μ con probabilidad 0.95.

Una vez que se ha extraído la muestra, el intervalo anterior deja de ser aleatorio, ya que los estimadores se sustituyen por sus valores en la muestra. Para la muestra extraída el intervalo de confianza que resulta es

$$\left(10 - 2.201\frac{4.1}{\sqrt{12}}\ ,\ 10 + 2.201\frac{4.1}{\sqrt{12}}\right).$$

□

Problema 3. Primera Semana. Se dispone de los siguientes datos referentes a 14 observaciones del par (X, Y):

$$\sum x_i = 517 \qquad \sum y_i = 346$$

$$\tfrac{1}{14}\sum x_i^2 = 2792.5 \qquad \tfrac{1}{14}\sum y_i^2 = 1246.7$$

$$\tfrac{1}{14}\sum x_i y_i = 1844.6$$

Se pide:

(a) Calcular la pendiente de la recta de regresión de Y sobre X.

(b) Obtener una medida del ajuste de dicha recta a la nube de puntos.

Datos auxiliares: Coeficiente de correlación $r = 0.98$

Solución

(a) La pendiente de la recta de regresión de Y sobre X viene dada por el cociente

$$\frac{cov_{x,y}}{v_x} = \frac{\frac{1}{n}\sum_{i=1}^n x_i y_i - \bar{x}\bar{y}}{\frac{1}{n}\sum_{i=1}^n x_i^2 - \bar{x}^2}$$

ó, alternativamente,

$$\frac{cov_{x,y}}{v_x} = r\sqrt{\frac{v_y}{v_x}}$$

Sustituyendo en la primera expresión queda

$$\frac{1844.6-\left(\frac{517}{14}\right)\left(\frac{346}{14}\right)}{2792.5-\left(\frac{517}{14}\right)^2}$$

(b) La varianza residual se define como el error cuadrático medio cometido con la recta de regresión de Y sobre X. Entonces, su valor nos dará una medida del ajuste de dicha recta a la nube de puntos. Con los datos del enunciado obtenemos

$$v_y=\frac{1}{n}\sum_{i=1}^{n}y_i^2-\bar{y}^2=1246.7-\left(\frac{346}{14}\right)^2=635.9$$

y

$$Varianza\ residual=635.9\left(1-0.98^2\right)=25.18$$

Sin embargo, si queremos una medida del error con mayor capacidad de interpretación, deberemos calcular el valor relativo dado por el cociente

$$\frac{Varianza\ residual}{v_y}=1-r^2$$

ó, equivalentemente, tomar el coeficiente r^2 como una medida del grado de ajuste: un valor próximo a 1 reflejará un buen ajuste y un valor cercano al 0 indicará la cualidad contraria. En definitiva, podríamos calcular directamente

$$r^2=0.96$$

concluyendo que, en este caso, el ajuste es bastante bueno.

□

Problema 4. Primera semana. Se piensa que el tiempo de respuesta de un equipo informático, cuando se le solicita cierto tipo de información, sigue una distribución exponencial con parámetro $\lambda=1$ seg. (por tanto, la densidad es $f(x)=e^{-x}$, para $x\geq 0$). Contraste dicha hipótesis utilizando los siguientes datos:

▶ Los puntos $0.22, 0.51, 0.92, 1.61$ determinan 5 clases equiprobables para la densidad citada.

▶ La frecuencia observada en cada una de estas clases, en un muestreo aleatorio de tamaño 40, es:

$$6, 8, 10, 7, 9$$

respectivamente.

Datos auxiliares: $\chi^2_{5;0.025}=12.833$; $\chi^2_{4;0.05}=9.488$; $\chi^2_{5;0.05}=11.07$

Solución

Comparamos las frecuencias observada y esperada mediante el estadístico $\chi^2 = \sum_{i=1}^{5} \frac{(O_i - e_i)^2}{e_i}$. Los resultados se muestran en la siguiente tabla

Clases	O_i	e_i	$(O_i - e_i)^2$	$\frac{(O_i-e_i)^2}{e_i}$	$\sum_{i=1}^{5} \frac{(O_i-e_i)^2}{e_i}$
Menores que 0.22	6	8	4	4/8	1.25
$0.22 - 0.51$	8	8	0	0	
$0.51 - 0.92$	10	8	4	4/8	
$0.92 - 1.61$	7	8	1	1/8	
Mayores que 1.61	9	8	1	1/8	

El procedimiento que vamos a aplicar consiste en rechazar la hipótesis de interés —para el nivel de significación α— si se satisface la desigualdad

$$\sum_{i=1}^{5} \frac{(O_i - e_i)^2}{e_i} > \chi^2_{5-1;\alpha}$$

En nuestro caso, el estadístico proporciona el valor 1.25 y, según los datos auxiliares, $\chi^2_{4;0.05} = 9.488$. Como consecuencia, no hay suficiente evidencia —al nivel de significación 0.05— contra la hipótesis de que "*el tiempo de respuesta sigue una distribución exponencial con parámetro* $\lambda = 1$ *seg.*"

□

Problema 1. Segunda Semana. Para estudiar la posible relación entre dos variables, X e Y, se registraron los valores de dichas variables en una muestra aleatoria formada por 20 individuos. De la información recogida se obtuvieron los siguientes datos estadísticos:

$$\frac{1}{20}\sum x_i = 53.69 \qquad \frac{1}{20}\sum y_i = 58.81$$

$$\frac{1}{20}\sum x_i^2 = 2948.82 \qquad \frac{1}{20}\sum y_i^2 = 3505.15$$

$$\frac{1}{20}\sum x_i y_i = 3197.39$$

Suponiendo un modelo de regresión lineal, ¿se puede concluir, con un nivel de significación $\alpha = 0.05$, que la pendiente de la recta de Y sobre X es positiva?

Datos auxiliares: $t_{18;0.05} = 1.734$, $t_{20;0.025} = 2.086$; *la estimación insesgada de la varianza* σ^2 *vale* 31.36 ; cuando $\beta_1 = 0$,

$$\frac{\hat{\beta}_1}{S_R\sqrt{\frac{1}{nv_x}}} \sim t_{n-2} \ .$$

Solución

La pregunta alude al contraste

$$H_0 : \beta_1 \leq 0 \quad , \quad H_1 : \beta_1 > 0$$

Entonces, al nivel $\alpha = 0.05$, rechazaremos la hipótesis nula —concluyendo que β_1 es positiva— cuando

$$\frac{\hat{\beta}_1}{S_R\sqrt{\frac{1}{20v_x}}} > t_{18;0.05}$$

Con los datos de este ejercicio resulta:

- $\hat{\beta}_1 = \frac{cov_{x,y}}{v_x} = \frac{3197.39-53.69\times 58.81}{2948.82-53.69^2} = 0.60$
- $S_R = \sqrt{31.36} = 5.6$
- $\sqrt{\frac{1}{20(2948.82-53.69^2)}} = 0.03$

y, como consecuencia,

$$\frac{\hat{\beta}_1}{S_R\sqrt{\frac{1}{nv_x}}} = \frac{0.6}{5.6\times 0.03} = 3.57$$

Dado que $t_{18;0.05} = 1.734$, los datos nos permiten concluir que la pendiente β_1 es positiva.

$\square$

Problema 2. Segunda semana. La distribución del error de medida de cierto aparato es una variable aleatoria con distribución normal de media y varianzas desconocidas. En 10 mediciones, tomadas al azar, se observaron los siguientes errores (expresados en la unidad de medida correspondiente).

8 10 11 10 12 10 9 8 13 9

(a) Obtener un intervalo de confianza, al 90 %, para el error de medida medio.

(b) Si la varianza del error de medida es $\sigma^2 = 2.5$ obtener un nuevo intervalo de confianza, con el mismo nivel de confianza, para la media del error de medida. Compárelo con el obtenido en el apartado anterior.

Solución

(a) Puesto que el error de medida sigue una distribución normal con varianza desconocida, la cantidad pivotal que se utiliza para obtener el intervalo de confianza es

$$U = \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$$

siendo

- μ la media del error de medida.
- $\bar{X}$ la media muestral.
- S^2 la cuasivarianza muestral.
- n el tamaño muestral.

En este caso, para la muestra extraída de tamaño $n = 10$, se tiene que

$$\bar{x} = \frac{8 + 10 + 11 + 10 + 12 + 10 + 9 + 8 + 13 + 9}{10} = 10$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{(8-10)^2 + (10-10)^2 + \cdots + (9-10)^2}{9} = 2.67$$

Emplearemos el mismo procedimiento que en el problema 2 de la primera semana. Ahora el tamaño muestral es 10 y la cantidad pivotal U tiene distribución t con 9 grados de libertad. Puesto que la probabilidad de que U esté comprendida entre los valores $-t_{9;0.05} = -1.833$ y $t_{9;0.05} = 1.833$, obtenidos de las tablas de la t_9, es 0.90 —área sombreada de la figura 2— se obtendrá que

$$0.90 = P(-1.833 \leq \frac{\bar{X} - \mu}{S/\sqrt{10}} \leq 1.833) = P(-1.833\frac{S}{\sqrt{10}} \leq \bar{X} - \mu \leq 1.833\frac{S}{\sqrt{10}}).$$

Despejando en la desigualdad anterior μ, resulta que

$$\left(\bar{X} - 1.833\frac{S}{\sqrt{10}}\ ,\ \bar{X} + 1.833\frac{S}{\sqrt{10}}\right)$$

es un intervalo aleatorio que cubre a μ con probabilidad 0.90.

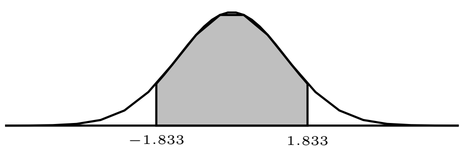


Figura 2: Función de densidad de una t_9

Para la muestra obtenida, el intervalo con nivel de confianza del 90 % será

$$\left(10 - 1.833\frac{\sqrt{2.67}}{\sqrt{10}}\ ,\ 10 + 1.833\frac{\sqrt{2.67}}{\sqrt{10}}\right).$$

(b) Puesto que ahora la varianza de la población es $\sigma^2 = 2.5$ conocida, no recurrimos, como hacíamos en el apartado anterior, a la distribución t, que elimina el efecto de la varianza. Utilizaremos la cantidad pivotal con distribución normal dada por

$$V = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

donde

- μ es la media del error de medida.
- $\bar{X}$ es la media muestral. Para la muestra extraída $\bar{x} = 10$.
- σ^2 es la varianza de la población. En este caso $\sigma^2 = 2.5$
- n es el tamaño muestral.

Al igual que en el apartado anterior, encontramos el intervalo que contiene a V con una probabilidad de 0.90. La distribución $N(0,1)$, que rige su comportamiento aleatorio, nos abre el camino. La probabilidad de que V, o equivalentemente, una $N(0,1)$, esté comprendida entre los valores $-z_{0.05} = -1.645$ y $z_{0.05} = 1.645$, obtenidos de las tablas de la normal, es 0.90 (área sombreada de la figura 3).

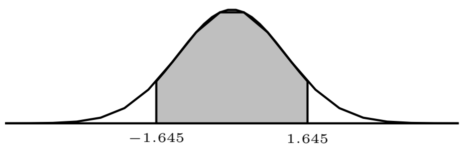


Figura 3: Función de densidad de una $N(0,1)$

Por tanto, se tendrá que

$$0.90 = P(-1.645 \leq \frac{\bar{X} - \mu}{\sqrt{2.5}/\sqrt{10}} \leq 1.645) = P(-1.645\frac{\sqrt{2.5}}{\sqrt{10}} \leq \bar{X} - \mu \leq 1.645\frac{\sqrt{2.5}}{\sqrt{10}})$$

Despejando μ en la desigualdad, resulta el siguiente intervalo de extremos aleatorios que cubre al error medio de medida con probabilidad 0.90

$$\left(\bar{X} - 1.645\frac{\sqrt{2.5}}{\sqrt{10}}\ ,\ \bar{X} + 1.645\frac{\sqrt{2.5}}{\sqrt{10}}\right)$$

Reemplazado el valor de la media muestral que resulta de la muestra extraída, obtendremos el siguiente intervalo con nivel de confianza del 90 %:

$$\left(10 - 1.645\frac{\sqrt{2.5}}{\sqrt{10}}\ ,\ 10 + 1.645\frac{\sqrt{2.5}}{\sqrt{10}}\right).$$

Puesto que para la muestra extraída $S^2 \approx \sigma^2$, las longitudes de ambos intervalos están caracterizadas por los cuantiles $t_{9;0.05}$ y $z_{0.05}$ de las distribuciones t y normal; el primero es mayor que el segundo ya que la distribución t tiene colas más pesadas que la normal. Esto explica que el intervalo de confianza basado en la t tenga mayor longitud que el basado en la normal.

□

Problema 3. Segunda semana. Un algoritmo de búsqueda inspecciona una lista de 1000 registros a fin de localizar un registro determinado. El algoritmo emplea un procedimiento secuencial de búsqueda: recorre la lista de izquierda a derecha, comprobando si cada registro coincide con el que busca, hasta que lo encuentra. Se pide:

(a) Calcular la probabilidad de que lo encuentre en 6 intentos.

(b) Calcular la probabilidad de que tenga que realizar k intentos.

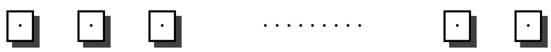
(c) Determinar el número medio de intentos que realiza.

Solución

(a) Sea X la variable aleatoria

$$X = \text{ número de intentos hasta encontrar el registro buscado.}$$

Supongamos que ponemos todos los registros en fila:



El algoritmo realizará seis intentos cuando no localice el registro que busca en las cinco primeras posiciones de la fila y lo encuentre en la sexta.

Si denotamos por A_i el suceso *“ el registro buscado ocupa la $i-$ésima posición de la fila ”* la probabilidad pedida será

$$P(X = 6) = P(A_1^c \cap A_2^c \cap A_3^c \cap A_4^c \cap A_5^c \cap A_6).$$

Por la regla de la multiplicación para el cálculo de la probabilidad de la intersección de sucesos (sección 6 del capítulo 3 del texto base), se tiene que

$$P(X=6) = P(A_1^c)P(A_2^c|A_1^c)P(A_3^c|A_1^c \cap A_2^c)\cdots P(A_6|A_1^c \cap A_2^c \cap A_3^c \cap A_4^c \cap A_5^c).$$

Por tanto, la probabilidad pedida será

$$P(X=6) = \left(\frac{999}{1000}\right)\left(\frac{998}{999}\right)\left(\frac{997}{998}\right)\left(\frac{996}{997}\right)\left(\frac{995}{996}\right)\left(\frac{1}{995}\right) = \frac{1}{1000}$$

(b) De la misma manera, la probabilidad de realizar k intentos es la probabilidad de que el algoritmo no localice el registro en los $k-1$ primeros lugares de la fila y lo encuentre en el k-ésimo. Por tanto, para cada $k = 1, 2, \ldots, 1000$

$$P(X=k) = \left(\frac{999}{1000}\right)\cdots\left(\frac{1000-k+1}{1000-k+2}\right)\left(\frac{1}{1000-k+1}\right) = \frac{1}{1000}$$

El cálculo anterior se generaliza sin dificultad a una lista con n registros. Así, para cada $k = 1, 2, \ldots, n$, se tiene que

$$P(X=k) = \left(\frac{n-1}{n}\right)\left(\frac{n-2}{n-1}\right)\cdots\left(\frac{n-k+1}{n-k+2}\right)\left(\frac{1}{n-k+1}\right) = \frac{1}{n}$$

Sin embargo, para calcular $P(X=k)$ en el caso general, preferimos utilizar el siguiente razonamiento que en matemáticas llamamos *recurrente* y que se sigue del procedimiento secuencial de búsqueda.

Denotaremos por $p_{k,n}$ la probabilidad de localizar el registro buscado en k intentos con una lista de n registros y por A el suceso *"el primer registro de la fila es distinto al buscado"*.

Para localizar el registro en k intentos, debe ocurrir A, y a continuación, se han de realizar $k-1$ intentos en una nueva lista con $n-1$ registros (todos menos el primero). Por tanto, la probabilidad del suceso $\{X=k\}$ será

$$p_{k,n} = P(X=k) = \frac{n-1}{n} p_{k-1,n-1} : k = 2, 3, \ldots, n$$

de donde se sigue la ecuación recurrente

$$np_{k,n} = (n-1)p_{k-1,n-1} \tag{1}$$

Teniendo en cuenta la condición inicial: $p_{1,i} = \frac{1}{i}$ (en una lista con i registros la probabilidad de localizar el buscado en el primer intento es $1/i$), basta aplicar la ecuación anterior sucesivamente para obtener que

$$np_{k,n} = (n-1)p_{k-1,n-1} = (n-2)p_{k-2,n-2} = \cdots\cdots = (n-k+1)p_{1,n-k+1} = 1$$

de donde se sigue que

$$p_{k,n} = P(X = k) = \frac{1}{n} : k = 1, 2, \ldots, n.$$

(c) Ya que hemos sido capaces de generalizar el problema, vamos a seguir utilizando la lista de n registros.

El número medio de intentos que realiza el algoritmo es la media de la variable aleatoria X.

$$E\{X\} = \sum_{k=1}^{n} kP(X = k) = \frac{1}{n}\sum_{k=1}^{n} k.$$

Calcular este sumatorio es un ejercicio sencillo si se tiene en cuenta que la suma de cada dos términos del sumatorio que equidistan de los sumandos extremos es igual a la suma de estos:

$$\begin{array}{ccccccc} 1 & 2 & 3 & \cdots\cdots\cdots & n-2 & n-1 & n \\ n & n-1 & n-2 & \cdots\cdots\cdots & 3 & 2 & 1 \\ \hline n+1 & n+1 & n+1 & \cdots\cdots\cdots & n+1 & n+1 & n+1 \end{array}$$

De lo anterior se sigue que $2\sum_{k=1}^{n} k = n(n+1)$, es decir,

$$E\{X\} = \frac{1}{n}\sum_{k=1}^{n} k = \frac{n+1}{2}.$$

Cuando $n = 1000$, el número medio de intentos es $1001/2$.

El razonamiento recurrente nos proporciona de nuevo un procedimiento de cálculo de la media que evita cuentas *"engorrosas"* como las anteriores.

Denotamos por μ_n el número medio de intentos en una lista con n registros.

Si el registro buscado está en la primera posición de la fila, lo cual ocurre con probabilidad $1/n$, se realiza un intento y se acaba la búsqueda. En cambio, si no está, lo cual ocurre con probabilidad $\frac{(n-1)}{n}$, contamos un intento y comenzaremos a buscar en una lista con $n-1$ registros; con lo que, en este caso, el número medio de intentos será $1 + \mu_{n-1}$.

De este razonamiento resulta la siguiente ecuación recurrente:

$$\mu_n = \frac{1}{n} + \frac{n-1}{n}(1 + \mu_{n-1}) \tag{2}$$

con la condición inicial $\mu_1 = 1$ (en una lista con un solo registro se localiza el buscado en un intento).

Poniendo $Q_n = n\mu_n$, la ecuación recurrente 2 se transforma en

$$Q_n = Q_{n-1} + n \qquad \text{con} \qquad Q_1 = \mu_1 = 1 \tag{3}$$

Es posible que no sepas resolver esta ecuación en diferencias. Realmente no lo necesitas, ya que el enunciado tan sólo te pide que encuentres $\mu_{1000} = \frac{Q_{1000}}{1000}$. Seguro que sí sabes programar un bucle que realice el cálculo. Las siguientes cinco líneas de código te resultarán familiares.

```
Q=1
for n = 2 to 1000
Q=Q+n
next n
Q/1000
```

Para los aficionados a resolver problemas, vamos a solucionar 3.

Ensayamos para Q_n una solución de la forma: $Q_n = a+bn+cn^2$. Partiendo de la condición inicial, basta aplicar la recurrencia dos veces para obtener

$$Q_1 = 1 \qquad Q_2 = 3 \qquad Q_3 = 6$$

Sustituyendo los valores $n = 1$, $n = 2$ y $n = 3$ en la solución general, se llega al siguiente sistema de ecuaciones.

$$\begin{array}{rcl} a + b + c & = & 1 \\ a + 2b + 4c & = & 3 \\ a + 3b + 9c & = & 6 \end{array}$$

La solución del sistema es $a = 0$, $b = 1/2$, $c = 1/2$; con lo que

$$\mu_n = \frac{Q_n}{n} = \frac{1}{n}\left(\frac{n}{2} + \frac{n^2}{2}\right) = \frac{n+1}{2}$$

□

Problema 4. Segunda Semana. A partir de una muestra de 26 observaciones de la variable X —que toma valores entre 320 y 430—, se obtuvo el siguiente diagrama de tallos y hojas:

32	55
33	49
34	
35	6699
36	34469
37	03345
38	9
39	2347
40	23
41	
42	4

(a) Reproduzca las 10 primeras observaciones (en la ordenación de menor a mayor).

(b) ¿Dónde está situada la mediana de la distribución? ¿Qué variación experimentaría dicha medida de centralización si el máximo de la distribución aumentara su valor en 10 unidades?

(c) Sabiendo que el valor medio es 370.7, ¿cómo mediría la dispersión de los datos respecto a este valor central? (No se requiere realizar los cálculos)

Solución

(a) Las observaciones pedidas son

$$325, 325, 334, 339, 356, 356, 359, 359, 363, 364$$

(b) La mediana de la distribución está situada en el punto

$$\frac{369 + 370}{2} = 369.5$$

Si el máximo de la distribución, que es 424, aumentara su valor en 10 unidades, la mediana estaría situada en el mismo punto —en 369.5—, ya que seguiríamos teniendo el mismo número de observaciones a cada lado.

(c) Mediante la desviación típica, que se define como la raíz cuadrada positiva de la varianza. Para el cálculo de esta última, se puede aplicar directamente la definición:

$$v_x = \frac{1}{n} \sum_i (x_i - \bar{x})^2$$

ó, equivalentemente,

$$\begin{aligned} v_x &= \frac{1}{n}\sum_i x_i^2 - \bar{x}^2 \\ &= \frac{1}{26}\left(325^2 + 325^2 + 334^2 + \ldots + 424^2\right) - 370.7^2 \end{aligned}$$

$\square$