

Tema 6. Descripción numérica (2)

Capítulo 5 del manual

Tema 6 – Descripción numérica (2)

Introducción

1. La mediana
 2. Los cuartiles
 3. El rango y el rango intercuartílico
 4. Los percentiles
 5. El diagrama de caja. Datos atípicos
 6. Comparación de media y mediana
 7. La media recortada
 8. Distribuciones de frecuencias
 9. La moda
 10. Estadística descriptiva con EXCEL
- Resumen
Ejercicios

Tema 6- Descripción numérica (2)

2

Introducción

- Tema anterior: descripción numérica basada en sumas de **valores**
- Ahora: descripción numérica basada en ordenaciones de **casos**
- Medidas de localización
- Medidas de dispersión
- Gráfico que reúne ambas cosas

Tema 6- Descripción numérica (2)

3

1. La mediana

- Dados unos casos de una variable:

$$x_1, x_2, \dots, x_{n-1}, x_n$$

- Esos casos, ordenados de menor a mayor se representan así:

$$x_{(1)}, x_{(2)}, \dots, x_{(n-2)}, x_{(n-1)}, x_{(n)}$$

- La mediana (med_x) es el punto (o puntos) que separa las observaciones ordenadas de menor a mayor en dos grupos con el mismo número de elementos.
- Es decir, el valor del caso central de la serie de datos, ordenados de menor a mayor

Tema 6- Descripción numérica (2)

4

1. La mediana: cálculo

- El valor “central” sería $x_{((n+1)/2)}$
- Si hay un número impar de observaciones: la mediana es el valor del caso central
 - ◆ Por ejemplo, con 45 datos: $x_{((45+1)/2)}=x_{(23)}$
- Con número par de casos: la mediana es un valor no existente realmente entre los casos observados
 - ◆ Por ejemplo, con 100 casos: $x_{((100+1)/2)}=x_{(50,5)}$
- ¿Cómo se calcula el valor del “caso 50,5” de una lista?
- Pues $x_{(50)}+0,5(x_{(51)}-x_{(50)})$
- Que es lo mismo que la media entre los valores de los casos 50 y 51

1. La mediana: cálculo para variables discretas

- La fórmula anterior se simplifica: la mediana será el valor que “incluya” la frecuencia relativa acumulada 0,5 (excepto si un valor tiene exactamente la frecuencia acumulada 0,5: entonces regla anterior)
- Ejemplo CAPITAS: el valor 4

Clase	Frecuencias		Frecuencias acumuladas	
	absolutas (n)	relativas (f)	Absolutas (N)	Relativas (F)
1	6	0,08	6	0,08
2	11	0,15	17	0,23
3	11	0,15	28	0,37
4	20	0,27	48	0,64
5	15	0,20	63	0,84
6	8	0,11	71	0,95
7	3	0,04	74	0,99
8	0	0,00	74	0,99
9	1	0,01	75	1,00
	75	1		

1. La mediana (4): Cálculo con EXCEL

- Cálculo con EXCEL
- Ejemplos de variables GTINE y AHORRO
- =mediana(rango)
- GTINE: 226.177 (libro dato diferente)
- AHRR: 0
- Comprobación
 - Marcar bloque, Datos, Ordenar, Columna (P ó O)
 - Buscar: observación número 38 $[(75+1)/2]$
 - OJO: TODAS LAS COLUMNAS CON DATOS

2. Los cuartiles

- Partiendo de la misma idea de la mediana (dividir los datos en bloques de igual número de datos) podemos introducir nuevas medidas de dispersión.
- Los **cuartiles**: valores que dividen el conjunto de observaciones, ordenadas de menor a mayor, en cuatro grupos con el mismo número de elementos.

2. Los cuartiles

- Si mediana divide el conjunto ordenado de casos en dos mitades:
 - ◆ el **primer cuartil** Q_1 , es la mediana de la mitad de los casos con los valores más pequeños
 - ◆ El **segundo cuartil** Q_2 , es la mediana
 - ◆ el **tercer cuartil** Q_3 , es la mediana de la mitad de los casos con los valores más grandes

2. Los cuartiles

- Otra forma de expresarlo:
 - ◆ El **primer cuartil** Q_1 es el menor valor que es mayor que el valor de una cuarta parte de los casos
 - ◆ El **segundo cuartil** Q_2 (la mediana), es el menor valor que es mayor que el valor de la mitad de los casos
 - ◆ El **tercer cuartil** Q_3 es el menor valor que es mayor que el valor de tres cuartas partes de los datos

2. Los cuartiles

- Aún otra forma de expresarlo (significado ligeramente distinto):
 - ◆ El **primer cuartil** Q_1 es el valor que tiene la frecuencia relativa acumulada 0,25
 - ◆ El **segundo cuartil** Q_2 (la mediana), es el valor que tiene la frecuencia relativa acumulada 0,50
 - ◆ El **tercer cuartil** Q_3 es el valor que tiene la frecuencia relativa acumulada 0,75

2. Los cuartiles: cálculo para variables discretas

- El **primer cuartil** Q_1 sería el primer valor que, en la distribución de frecuencias, supera (o iguala) la frecuencia relativa acumulada 0,25. En este caso: el valor 3
- El **tercer cuartil** Q_3 sería el primer valor que, en la distribución de frecuencias, supera (o iguala) la frecuencia relativa acumulada 0,75. En este caso: el valor 5

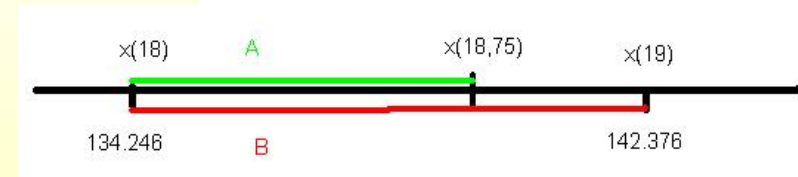
Variable CAPITAS (número miembros hogar)				
Clase	Frecuencias		Frecuencias acumuladas	
	absolutas (n)	relativas (f)	Absolutas (N)	Relativas (F)
1	6	0,08	6	0,08
2	11	0,15	17	0,23
3	11	0,15	28	0,37
4	20	0,27	48	0,64
5	15	0,20	63	0,84
6	8	0,11	71	0,95
7	3	0,04	74	0,99
8	0	0,00	74	0,99
9	1	0,01	75	1,00
	75	1		

2. Los cuartiles : cálculo para variables continuas

- El **primer cuartil** Q_1 sería el valor que tenga la frecuencia relativa acumulada 0,25
- Normalmente no habrá ningún valor en la muestra o población que tenga exactamente esa frecuencia relativa acumulada
- Cálculo por “interpolación”
- Ejemplo: en la variable GTINE, que tiene 75 casos, el caso con la frecuencia acumulada 0,25, sería el caso $x_{(75/4)} = x_{(18,75)}$
- Pero el caso 18,75 no existe; existen los casos 18 y 19 (recordemos: ordenados de menor a mayor)

2. Los cuartiles : cálculo para variables continuas

- ¿Cuál sería el valor de un hipotético caso 18,75? Pues el valor del caso 18 más 0,75 multiplicado por la diferencia entre el valor 18 y el valor 19
- $x_{(18,75)} = x_{(18)} + [0,75*(x_{(19)}-x_{(18)})]$



2. Los cuartiles: cálculo para variables continuas

- Por ejemplo, para GTINE, si ordenamos los casos de menor a mayor valor, estos serían los dos casos $x_{(18)}$ y $x_{(19)}$

Caso número	Valor	Frec. Relativa acumulada
18	134.246	0,240
19	142.376	0,253

- Entonces $Q_1 = x_{(18,75)} = x_{(18)} + [0,75*(x_{(19)}-x_{(18)})]$
- $Q_1 = 134.246 + [0,75*(142.376-134.246)] = 134.246 + [0,75*8.130] = 134.246 + 6.097,5 = 140.343,5$

2. Los cuartiles: cálculo para variables continuas

- El **tercer cuartil** Q_3 sería el valor que tenga la frecuencia relativa acumulada 0,75
- Volvemos a calcular por “interpolación”
- El valor que buscamos sería el valor $x_{(75*0,75)} = x_{(56,25)}$
- Ese valor no existe, pero lo podemos calcular como

$$Q_3 = x_{(56,25)} = x_{(56)} + [0,25*(x_{(57)}-x_{(56)})]$$

2. Los cuartiles: cálculo para variables continuas

- Por ejemplo, para GTINE, si ordenamos los casos de menor a mayor valor, estos serían los dos casos $x_{(56)}$ y $x_{(57)}$

Caso número	Valor	Frec. Relativa acumulada
56	309.787	0,747
57	309.964	0,760

- $Q_3 = x_{(56,25)} = x_{(56)} + [0,25*(x_{(57)} - x_{(56)})]$
- Entonces $Q_3 = 309.787 + [0,25*(309.964 - 309.787)] = 309.787 + [0,25*177] = 309.787 + 44,25 = 309.831,25$

2. Los cuartiles :con EXCEL

Función CUARTIL devuelve los siguientes valores:

- =CUARTIL(rango;0) da el valor mínimo
 - =CUARTIL(rango;1) da el primer cuartil
 - =CUARTIL(rango;2) da la mediana
 - =CUARTIL(rango;3) da el tercer cuartil
 - =CUARTIL(rango;4) da el valor máximo
- Ejemplo con GTINE
 - 48.586, 142.527, 226.177, 309.875, 876.161
 - Q_1 y Q_3 son diferentes a los calculados por nosotros
 - “Anomalía” de Excel, sobre todo si número de casos es pequeño
 - OJO: CUIDADO CON CUARTILES SI POCOS CASOS

3. El rango y el rango intercuartílico

- Medida elemental de dispersión: el rango R , que es la diferencia entre el valor mayor y el más pequeño

$$R_x = x_{(N)} - x_{(1)}$$

- Con EXCEL: =MAX(rango)-MIN(rango)

3. El rango y el rango intercuartílico

- El **rango intercuartílico** es la diferencia entre el tercer y el primer cuartil:

$$RI_x = Q_{3x} - Q_{1x}$$

- Con EXCEL:
=CUARTIL(rango;3)-CUARTIL(rango;1)

4. Los percentiles

- Generalizando el concepto de cuartiles: percentiles
 - ◆ Si primer cuartil es el dato más pequeño que es mayor que un cuarto de los datos
 - ◆ Y tercer cuartil es el dato más pequeño que es mayor que tres cuartos de los datos
 - ◆ El **percentil de orden p** es el dato más pequeño que es mayor que el **p** por ciento de los datos

4. Los percentiles

- El percentil 10 P_{10} sería el valor que tenga la frecuencia relativa acumulada 0,10
- Si son variables discretas: hacemos como hemos visto antes: el percentil 10 será el valor cuya frecuencia acumulada “contenga” el 0,10
- Por ejemplo, en transparencia 12: ¿qué valor es el percentil 10? ¿Y el percentil 90?

4. Los percentiles

- Con variables continuas: interpolación, otra vez
- Por ejemplo, para GTINE, si ordenamos los casos de menor a mayor valor, estos serían los dos casos con F_i menor y mayor que 0,10

Caso número	Valor	Frecuencia acumulada
7	81.861	0,093
8	87.112	0,107

- $P_{10} = x_{(75*0,10)} = x_{(7,5)} = x_{(7)} + [0,5*(x_{(8)} - x_{(7)})]$
- Entonces $P_{10} = 81.861 + [0,5*(87.112 - 81.861)] = 81.861 + [0,5*5.251] = 81.861 + 2.625,55 = 84.486,5$

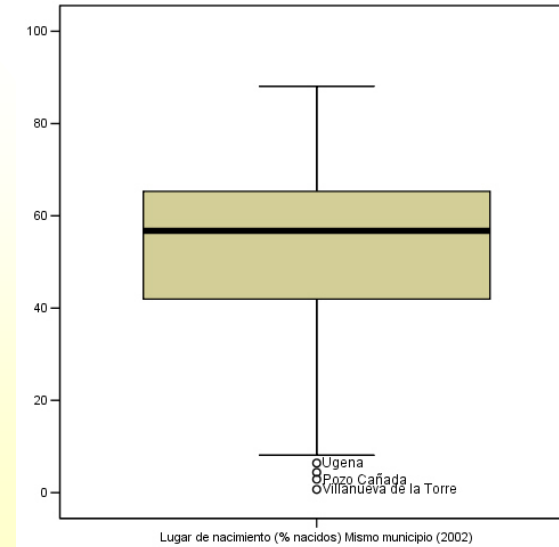
4. Los percentiles: con EXCEL

- Con EXCEL, usamos la función
 - =PERCENTIL(RANGO; valor del percentil)
- Valor del percentil: expresado entre 0 y 1
- Por ejemplo: percentil 10
 - =PERCENTIL(GTINE;0,10)
 - 88.066
- Percentil 90: =PERCENTIL(GTINE;0,9)
- Resultado: 472.201
- Como cuartiles: resultados no idénticos

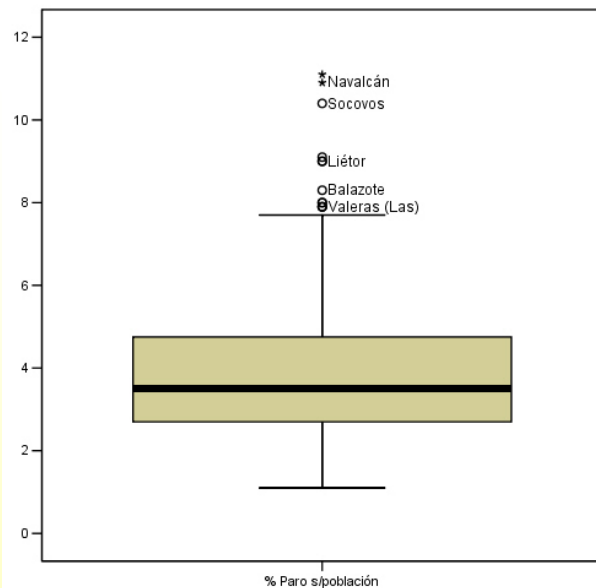
5. El diagrama de caja

- Gráfico basado en los cuartiles
- También informa sobre asimetría de distribución
- OJO: libro en horizontal; aquí en vertical
- Qué representa:
 - ◆ Caja: límite inferior y superior son Q_1 y Q_3 (RI)
 - ◆ Línea horizontal interior: mediana
 - ◆ Línea superior e inferior: unen caja con valor más alto y más bajo, hasta $(Q_1 - 1,5 RI)$ y $(Q_3 + 1,5 RI)$
 - ◆ Puntos: **datos atípicos** (*outliers*): desde 1,5 a 3 RI
 - ◆ Cruces: **datos atípicos extremos**: más de 3 RI

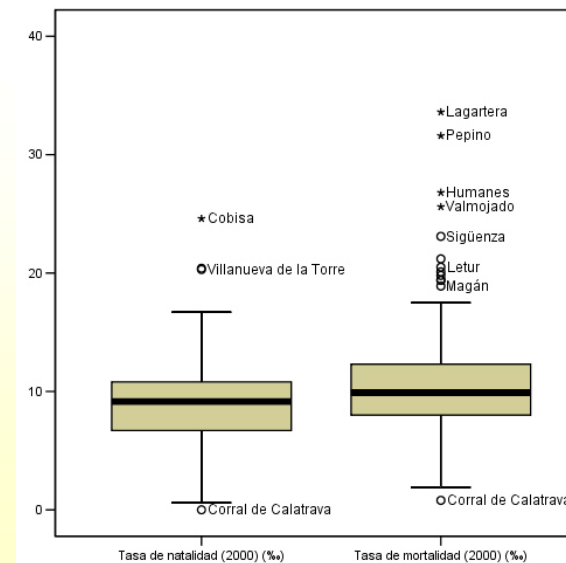
5. El diagrama de caja: ejemplos



5. El diagrama de caja: ejemplos



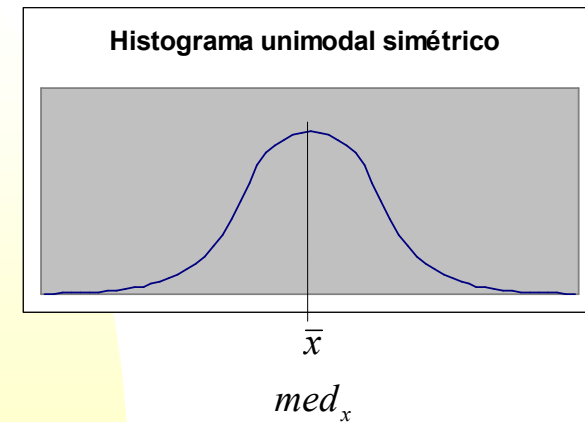
5. El diagrama de caja: ejemplos



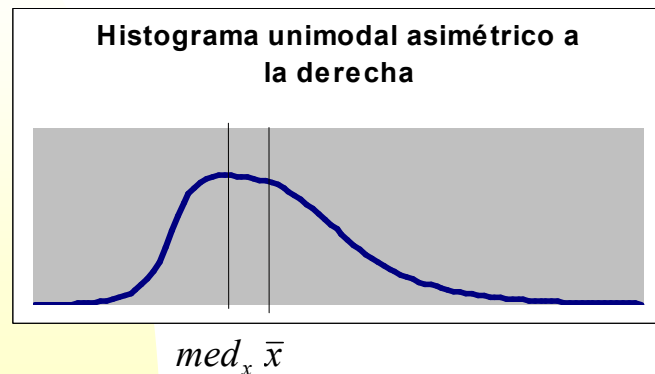
6. Comparación de mediana y media: robustez

- Media y mediana muy distintas ante datos atípicos
- Ejemplo: si en serie
1,7 2,8 3,2 3,4 5,3 5,9 6,2 7,2 8,3 9,3
Media= 5,3
Mediana= 5,6
- Por error 8,3 se tecldea como 83
Media= 12,8
Mediana= 5,6

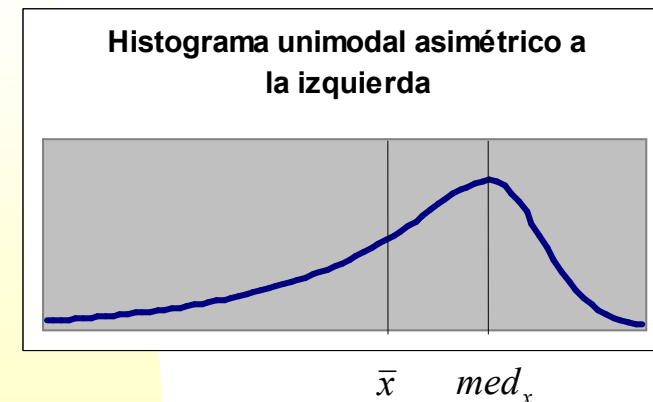
6. Comparación de mediana y media: distribuciones simétricas



6. Comparación de mediana y media: distribuciones asimétricas a la derecha



6. Comparación de mediana y media: distribuciones asimétricas a la izquierda



6. Comparación de mediana y media: conclusiones

- Media y desviación típica recomendadas para distribuciones homogéneas (simétricas y sin datos atípicos)
- Mediana y rango intercuartílico preferible para otros datos

7. La media recortada

- Posible remedio al problema de la media con los valores atípicos (falta de robustez)
- Supresión de los valores más extremos
- Media recortada al α por ciento es la media de los datos que quedan después de eliminar el α por ciento de los datos más grandes y el α por ciento de los datos más pequeños

7. La media recortada: ejemplo

- Valores de ejemplo anterior: (con error)
1,7 2,8 3,2 3,4 5,3 5,9 6,2 7,2 9,3 83
- Media (con error): 12,8
- Media recortada al 10 por ciento:
 - ◆ Son diez valores: el 10 por ciento es uno
 - ◆ Quitar valor menor y mayor
 - ◆ Media recortada: 5,41
- Media recortada al α por ciento elimina el efecto de los datos atípicos si la proporción de ellos en cada extremo es menor que α

8. La distribución de frecuencias y las medidas basadas en el orden

- ¿Qué hacer si tenemos sólo distribución de frecuencias?
- Si variables discretas: ningún problema (ya lo hemos usado)
- Variables continuas: ver las frecuencias acumuladas y tomar la marca de clase para calcular
 - ◆ Mediana: el valor en el que frecuencias acumuladas pasan la marca del 0,5
 - ◆ Cuartiles: valores que “contienen” la F_i 0,25 y 0,75
 - ◆ Percentiles: valores que “contienen” la F_i correspondiente
 - ◆ Rango intercuartílico: resta de $Q_3 - Q_1$

8. La distribución de frecuencias: variables continuas

Distribución de frecuencias variable GTINE				
Marca de clase	Frecuencias		Frecuencias acumuladas	
	Absolutas <i>n</i>	Relativas <i>f</i>	Absolutas <i>N</i>	Relativas <i>F</i>
25.000	1	0,01	1	0,01
75.000	10	0,13	11	0,15
125.000	9	0,12	20	0,27
175.000	12	0,16	32	0,43
225.000	11	0,15	43	0,57
275.000	11	0,15	54	0,72
325.000	3	0,04	57	0,76
375.000	1	0,01	58	0,77
425.000	6	0,08	64	0,85
475.000	5	0,07	69	0,92
525.000	1	0,01	70	0,93
575.000	0	0,00	70	0,93
625.000	2	0,03	72	0,96
675.000	1	0,01	73	0,97
725.000	1	0,01	74	0,99
775.000	0	0,00	74	0,99
825.000	0	0,00	74	0,99
875.000	1	0,01	75	1,00
	75	1,00		

Tema 6- Descripción numérica (2)

37

8. La distribución de frecuencias y las medidas basadas en el orden: variables continuas

- Cálculo con distribución de frecuencias, comparado con cálculo con datos originales:

GTINE	Distrib. Frecuencias	Datos originales
Mediana	225.000	226.177 (trans.7)
Primer cuartil	125.000	140.343,5 (trans. 15)
Tercer cuartil	325.000	309.831,25 (trans. 17)
Rango Intercuartílico	200.000	167.588

Tema 6- Descripción numérica (2)

38

9. La moda

- La media o la mediana: medidas de valor “central” en variables cuantitativas
- ¿Y en variables cualitativas? ¿Cómo resumir con un solo valor?
- La moda: el dato o clase o categoría más frecuente (el que tiene mayor frecuencia absoluta o relativa)
- Ejemplo: SITPROF
La moda es “empresarios sin empleados”
- OJO: no confundir la moda con la frecuencia relativa de la moda
- La moda puede usarse también para variables cuantitativas
- Dificultad: varios valores con frecuencias similares

Tema 6- Descripción numérica (2)

39

10. Estadística descriptiva con EXCEL

- Hasta ahora: estadísticas con EXCEL, una a una
- Procedimiento “expres” para ver el resumen en estadística descriptiva de una variable
- Herramientas – Análisis de datos – Estadística descriptiva
 - Indicar el rango
 - Marcar al menos “Resumen de estadísticas”

Tema 6- Descripción numérica (2)

10. Estadística descriptiva con EXCEL

- Obtenemos una tabla como esta (gtine):

Media	261542,04
Error típico	19738,59428
Mediana	226177
Moda	#N/A
Desviación estándar	170941,2408
Varianza de la muestra	29220907816
Curtosis	1,805270885
Coefficiente de asimetría	1,325257024
Rango	827575
Mínimo	48586
Máximo	876161
Suma	19615653
Cuenta	75

Tema 6- Descripción numérica (2)

10. Estadística descriptiva con EXCEL

- Nos podemos olvidar del error típico (volveremos sobre él en la parte de la inferencia)
- También nos olvidamos aquí (el libro lo explica) de la curtosis
- Todos los demás: los conocemos
- Faltan algunos que hemos visto (coeficiente de variación, cuartiles y rango intercuartílico)

Tema 6- Descripción numérica (2)

Resumen

- Conceptos y formas de determinar:
 - ◆ La mediana
 - ◆ Los cuartiles
 - ◆ El rango
 - ◆ El rango intercuartílico
 - ◆ Los percentiles
 - ◆ El diagrama de caja
 - ◆ La media recortada
 - ◆ Para variables discretas y continuas
 - ◆ Todos ellos también para distribuciones de frecuencias
 - ◆ La moda

Tema 6- Descripción numérica (2)

43

Ejercicios recomendados

- Del manual:
 - ◆ 5.1 (excepto c))
 - ◆ 5.2 (excepto c))
 - ◆ 5.3 (excepto b))
 - ◆ 5.7 (a, b, c)
 - ◆ 5.10 (excepto d y f)
 - ◆ Las excepciones se refieren a dibujar diagramas de caja. Excel no lo hace, pero podéis intentarlo, aproximadamente, “a mano”

Tema 6- Descripción numérica (2)

44

Ejercicios recomendados

■ De exámenes:

- ◆ Feb02: 2b, 3eg, 8
- ◆ Jun02: 2b, 3eg, 5
- ◆ Feb03: 2b, 3e, 4
- ◆ Sep03: 2e, 3, 4-2
- ◆ Feb04: 3b, 6, 7
- ◆ Jul04: 3b, 5ef, 6
- ◆ Feb05, Jun05: 5, 6
- ◆ Feb06: 3defg, 4a,
- ◆ Jun06: 3defg, 4ab,
- ◆ Ene07: 3defg
- ◆ Jul07: 3defg, 4
- ◆ Ene08, Jun08: 3defg,