



UNIVERSIDADE DA CORUÑA



UNIVERSIDADE
DE VIGO

Máster en Técnicas Estadísticas

Memoria de las prácticas realizadas en la unidad de epidemiología clínica del CHUS

Fiz Lagoa Labrador

Julio 2012

Resumen

En este informe de prácticas realizadas en la Unidad Epidemiológica del Complejo Hospitalario de Santiago de Compostela (CHUS), se resume en cuatro capítulos las tareas realizadas por el alumno Fiz Lagoa Labrador. Además de estas tareas, dicho alumno participó en los estadillos que periódicamente se realizan en esta unidad, tanto de oyente como de ponente (exponiendo algunos temas como Series de Tiempo y Contrastes de simetría en tablas cuadradas).

En el Capítulo N°1: *Quejas en un hospital universitario*, se realizará un estudio descriptivo y una evolución temporal de las quejas dependiendo del año, de los meses, del motivo y del servicio clínico.

En el Capítulo N°2: *Concentración de elastasa leucocitaria en la enfermedad arterial coronaria diagnosticada mediante angiografía*, se mostrará las características clínicas de los pacientes del estudio y de los que tengan enfermedad arterial coronaria. También se analizará la asociación de la concentración de elastasa y la enfermedad arterial coronaria y se evaluará la capacidad diagnóstica de la elastasa en dicha enfermedad.

En el Capítulo N°3: *Funciones en R para un análisis descriptivo y comparativo de las variables cualitativas y cuantitativas*, se construirá tres funciones para describir y comparar: dos variables cualitativas, dos variables cuantitativas y una variable cualitativa y otra cuantitativa.

En el Capítulo N°4: *Evaluación del impacto de una consulta especializada en el servicio de medicina interna*, se caracterizarán a los individuos controlados en la consulta especializada frente a los no controlados, se caracterizarán a los individuos controlados según el motivo de finalización del seguimiento en la consulta especializada y se definirán los factores pronóstico de estos motivos, se describirán las causas de muerte (cardiovasculares y no cardiovasculares) y se definirán los factores pronóstico de mortalidad, se describirá la repercusión del control de los pacientes hasta la fecha fin de seguimiento y finalmente se valorará la supervivencia de los pacientes con insuficiencia cardíaca controlados durante un periodo de 4 años y medio.

A continuación mostramos una tabla resumen en la cual aparecen los diferentes análisis o técnicas estadísticas utilizadas en cada capítulo y los paquetes en R que son utilizados.

Tabla Resumen

Capítulo	Técnicas Estadísticas	Paquetes
Capítulo N°1	Diagramas de Pareto Series de Tiempo	qcc
Capítulo N°2	Test de Lilliefors Test de Shapiro-Wilk Test chi-cuadrado Test exacto de Fisher Prueba t de comparación de medias Prueba F de comparación de varianzas Test de Wilcoxon Regresión Logística Curvas ROC	nortest DiagnosisMed
Capítulo N°3	Test de Lilliefors Test de Shapiro-Wilk Test chi-cuadrado Test Exacto de Fisher Test de Mc-Nemar Extensión de Bowker al test de Mc-Nemar Extensión de Stuart y Maxwell al test de Mc-Nemar Prueba t de comparación de medias Prueba F de comparación de varianzas Test de Wilcoxon Anova de un factor Test de Levene Método de Tukey Test de Kruskal-Wallis Test de Friedman	nortest coin car
Capítulo N°4	Test de Lilliefors Test de Shapiro-Wilk Test chi-cuadrado Test Exacto de Fisher Test de Mc-Nemar Prueba t de comparación de medias Prueba F de comparación de varianzas Anova de un factor Test de Levene Test de Wilcoxon Test de Kruskal-Wallis Regresión Logística Regresión Multinomial Estimador de Kaplan-Meier Test de Mantel y Haenzel Resgresión de Cox	nortest coin car effects multinom nnet survival

Índice

Capítulo 1:

<i>Quejas en un hospital universitario: estudio descriptivo y evolución temporal</i>	7
1. Introducción.....	7
2. Métodos	7
3. Resultados.....	9
3.1. Evolución anual	9
3.2. Evolución mensual	10
3.3. Motivos.....	11
3.4. Servicios clínicos.....	13
3.5. Distribución de los diferentes motivos	16
4. Conclusiones.....	17

Capítulo2:

<i>Concentración de elastasa leucocitaria en la enfermedad arterial coronaria diagnosticada mediante angiografía</i>	19
1. Introducción.....	19
2. Métodos	19
3. Resultados.....	20
3.1. Características clínicas de los pacientes del estudio.....	20
3.2. Características clínicas del grupo CAD.....	21
3.3. Análisis de la asociación entre la concentración de la elastasa y la enfermedad arterial coronaria (CAD).....	22
3.4. Capacidad diagnóstica de un test.....	23
4. Conclusiones.....	28

Capítulo3:

<i>Funciones en R para un análisis descriptivo y comparativo de variables cualitativas y cuantitativas</i>	29
1. Introducción.....	29
2. Funciones en R para un análisis descriptivo de los datos.....	30
3. Funciones en R para un análisis comparativo de los datos	30
3.1. Función en R para un análisis comparativo de dos variables cualitativas	31
3.2. Función en R para un análisis comparativo de dos variables cuantitativas	33

3.3. Función en R para un análisis comparativo de una variable cuantitativa y otra cualitativa	33
4. Ejemplo del uso de las funciones	34

Capítulo4:

Evaluación del impacto de una consulta especializada en el servicio de medicina interna.. 39

1. Introducción.....	39
2. Métodos	40
2.1. Diseño del estudio.....	40
2.2. Sujetos.....	40
2.3. Variables	40
3. Análisis estadístico	43
3.1. Test paramétricos en dos poblaciones normales.....	44
3.2. Test paramétrico en más de dos poblaciones normales	45
3.3. Test no paramétricos en dos muestras	48
3.4. Test no paramétrico para más de dos muestras.....	49
3.5. Independencia en tablas de contingencia.....	50
3.6. Regresión Logística	56
3.7. Regresión Multinomial	57
3.8. Análisis de Supervivencia.....	58
4. Resultados.....	65
4.1. Características clínicas de los individuos según el control.....	65
4.2. Características clínicas de los individuos según el status	66
4.3. Características clínicas de los individuos controlados según los motivos de finalización del seguimiento	68
4.4. Repercusión del seguimiento	70
4.5. Factores pronóstico de la mortalidad	72
4.6. Factores pronóstico de los motivos finalización del seguimiento	73
4.7. Valoración de la supervivencia.....	76
5. Conclusiones.....	82
Valoración de las prácticas	83
Bibliografía.....	85
Anexo	87

Capítulo 1

Quejas en un Hospital Universitario: Estudio descriptivo y evolución temporal

1. INTRODUCCIÓN

Existe una necesidad creciente en la mejora de la calidad en la provisión de cuidados de salud. Pacientes y familiares son la mejor fuente de datos para obtener información acerca de la calidad en la atención recibida.

La exploración y monitorización de las quejas que presentan los pacientes pueden ser buenos indicadores de la calidad de los cuidados y puede contribuir al diseño de estrategias que permitan mejorar la atención a los pacientes.

El objetivo de este estudio es describir los motivos y principales servicios de procedencia de las quejas presentadas en el periodo 2001-2011 por los pacientes o sus familiares.

2. MÉTODOS

Ámbito

Estudio retrospectivo de todas las quejas registradas por escrito en un Hospital Universitario. El periodo de estudio comprende desde Enero de 2001 hasta Octubre de 2011.

Variables

Se recoge la fecha de reclamación, domicilio del afectado, motivo de la reclamación, servicio hospitalario involucrado.

Las quejas han sido agrupadas en 8 motivos:

- *Deficiencia en la información*: Recoge deficiencia en la información en los procedimientos administrativos y procedimientos clínicos, información deficiente en el servicio de urgencias, ausencia de confidencialidad, información deficiente en el consentimiento informado, así como deficiencias en la señalización.

- *Deficiencia en la documentación:* No localización de la historia clínica, pérdida de documentación, retraso en la entrega de documentación clínica u otro tipo de documentación, así como no facilitar receta médica.
- *Lista de espera:* Se refiere a la lista de espera de cirugía.
- *Demora lista de espera:* Agrupa retrasos en las citas a consultas, realización de pruebas, suspensión de pruebas, no citaciones por falta de agenda y errores administrativos en la citación.
- *Demora en la atención:* Agrupa la no atención por falta de historia clínica, retraso en la atención de la cita, retraso en la atención en urgencias y quejas por el orden de entrada.
- *Desacuerdo en el proceso asistencial:* Agrupa desacuerdos con el diagnóstico y tratamiento, cambios de médicos y enfermería, traslados, altas, descoordinación entre servicios, suspensión o retraso de procedimientos quirúrgicos, yatrogenia, errores en el diagnóstico y tratamiento.
- *Trato personal:* Agrupa lo que el paciente o familiares consideran trato incorrecto o irrespetuoso, no identificación de profesionales, relación con otros pacientes y desacuerdo con el idioma utilizado.
- *Hostelería, habitabilidad y confort:* Agrupa deficiencias en limpieza, problemas con instalaciones, deficiencia con las comidas, fumar en zonas prohibidas, ruidos, temperatura inadecuada, problemas de aparcamiento, cafetería y temas relacionados con el confort en general.

Se han elegido los 10 servicios que presentan una mayor frecuentación de quejas, debido a la confidencialidad de los datos, a estos servicios los llamaremos: *servicio A, servicio B, servicio C, servicio D, servicio E, servicio F, servicio G, servicio H, servicio I y servicio J* para hacer un análisis específico de estos servicios.

Análisis estadístico

La descripción de los datos se presenta mediante tablas con datos brutos y porcentajes. Irán acompañados de Diagramas de Pareto en los cuales se puede observar mediante barras el número bruto de datos, su porcentaje y el porcentaje acumulado. Se realiza también un análisis de la evolución temporal del número de quejas en general así como para los servicios con mayor número de quejas.

Se presentarán también análisis descriptivos para los motivos agrupados según las definiciones arriba mencionadas.

3. RESULTADOS

3.1. Evolución anual

Entre los años 2001 y 2010 se han recogido 33521 quejas. En la Figura.1 mostramos un diagrama de Pareto, el cual muestra la frecuencia absoluta de quejas por año y el porcentaje acumulado, también un gráfico con la evolución temporal de dichas quejas. En la Tabla.1 se muestra el número total de quejas por año y sus porcentajes. Observamos como el número de quejas aumentan año a año, siendo los años 2006 y 2007 los años con mayor número de quejas. A partir del 2007 el número de quejas comienza a descender y en los años 2009 y 2010 vuelven a situarse al nivel del 2004-2005.

Figura.1 Diagrama de Pareto por Años y Gráfico de la evolución anual de las quejas

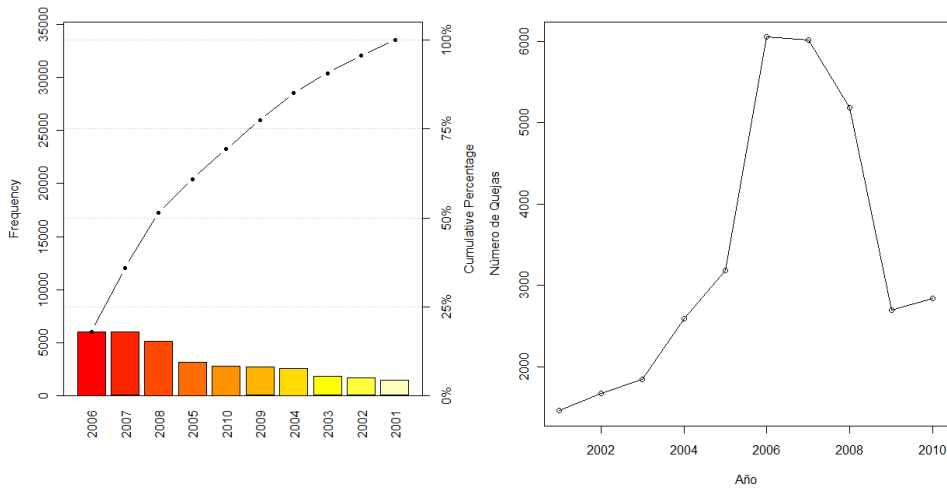


Tabla.1 Distribución anual de las quejas

Año	Número de quejas	Porcentajes
2001	1456	4.34%
2002	1666	4.97%
2003	1846	5.51%
2004	2591	7.73%
2005	3186	9.50%
2006	6050	18.05%
2007	6015	17.95%
2008	5179	15.45%
2009	2697	8.05%
2010	2835	8.46%
Total	33521	100%

3.2. Evolución mensual

La distribución de las quejas dentro del año puede no ser uniforme y puede haber una mayor frecuencia en algunos meses que en otros. En la Tabla.2 se presenta el número bruto de quejas distribuidas por mes y año.

Tabla.2 Distribución mensual de las quejas en cada año

	Ene	Feb	Mar	Abr	May	Jun	Jul	Ago	Sept	Oct	Nov	Dic
2001	126	134	196	103	98	106	109	89	109	115	156	115
2002	146	137	141	154	196	176	117	90	178	102	123	106
2003	104	127	146	122	140	183	177	68	169	199	209	202
2004	127	195	198	191	156	241	218	190	259	308	243	265
2005	161	234	180	316	266	335	301	132	194	372	337	358
2006	513	353	509	486	517	615	671	495	457	486	507	441
2007	363	638	558	437	578	773	422	369	440	530	526	381
2008	410	494	586	694	461	534	367	266	349	458	328	232
2009	226	210	319	270	284	267	198	149	152	170	213	239
2010	145	196	250	270	263	304	205	155	255	273	308	211
2011	235	277	277	187	208	227	212	180	204	244		

En la Figura.2 y la Tabla.3 se muestra la distribución de las quejas y porcentajes por meses. Se observa que los meses con mayor frecuencia de quejas son Junio y Marzo, los cuales tienen más de un 10% y 9% del total de las quejas respectivamente. Los meses con menor frecuencia son Enero y Agosto con menos del 13% de las quejas entre los dos meses. A la vista de estos datos puede observarse que en los meses de primavera y otoño son en los que se producen mayor número de quejas, mientras que los meses de invierno y verano disminuyen las quejas. En la Figura.2 también mostramos la evolución de las quejas en todo el periodo de estudio mes a mes.

Figura.2 Diagrama de Pareto por Meses y Gráfico de la evolución mensual de las quejas

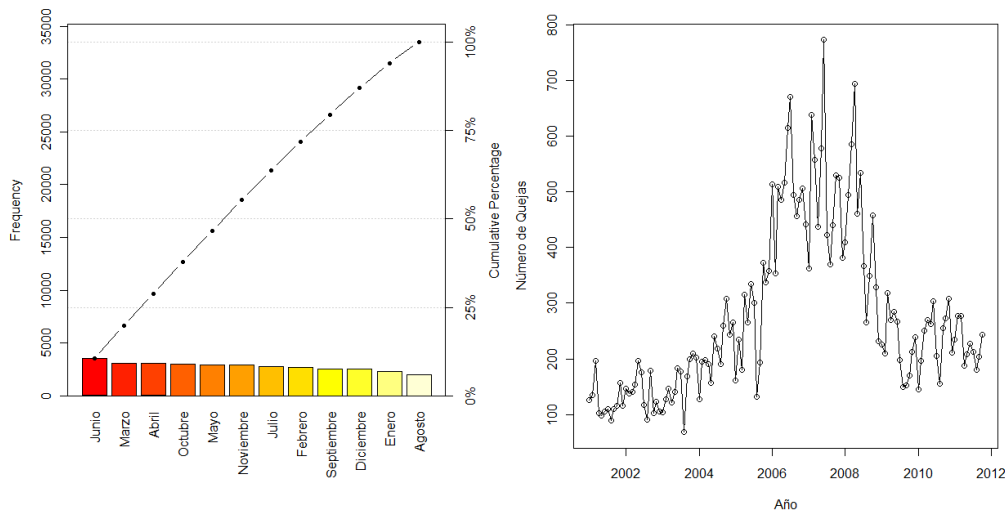


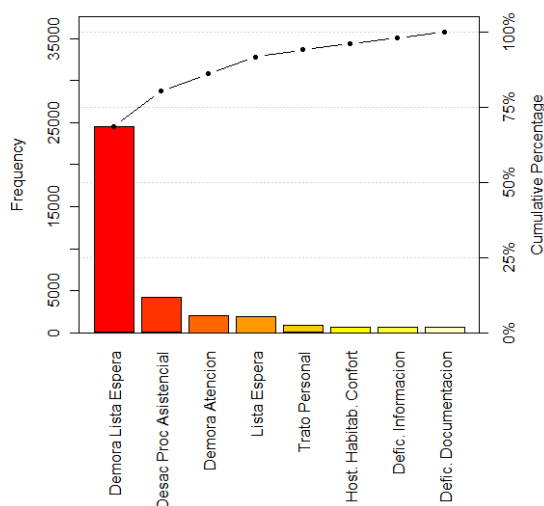
Tabla.3 Distribución mensual de las quejas

Año	Número de quejas	Porcentajes
Enero	2321	6.92%
Febrero	2718	8.11%
Marzo	3083	9.20%
Abril	3043	9.08%
Mayo	2959	8.83%
Junio	3534	10.54%
Julio	2785	8.31%
Agosto	2003	5.98%
Septiembre	2562	7.64%
Octubre	3013	8.99%
Noviembre	2950	8.80%
Diciembre	2550	7.61%
Total	33521	100%

3.3. Motivos

Como ya hemos comentado también se mostrará los resultados del análisis descriptivo para las quejas por los motivos de las mismas. Estos resultados son del periodo comprendido entre Enero del 2001 hasta Octubre del 2011. Así, en la Figura.3 mostramos un Diagrama de Pareto donde se muestran el número de quejas según los 8 grupos en que hemos categorizado los motivos.

Figura.3 Diagrama de Pareto por motivos



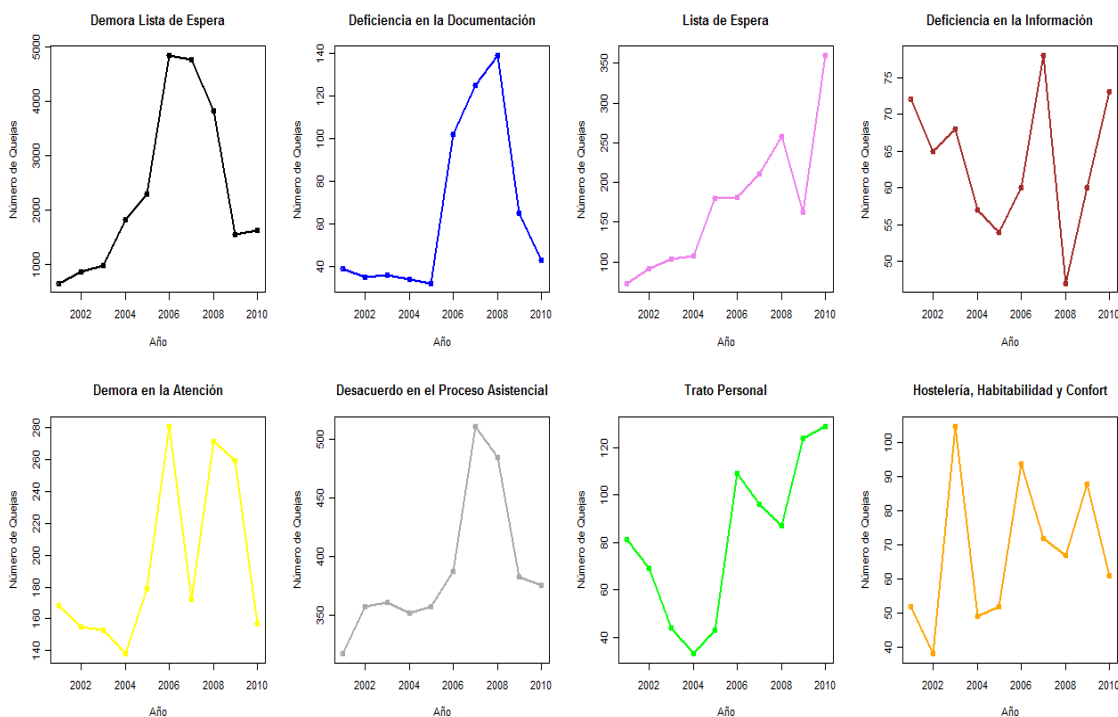
Como observamos en la Tabla.4, dos terceras partes de las quejas se producen en el grupo de quejas *Demora en la lista de espera*, teniendo el resto de grupos de quejas un porcentaje sobre el total muy pequeño. En la Tabla.4 mostramos el número total de quejas por cada motivo y sus respectivos porcentajes exactos.

Tabla.4 Distribución de las quejas por motivos

Motivos	Número de quejas	Porcentaje
Demora lista espera	24511	68.52%
Desacuerdo proceso asistencial	4198	11.73%
Demora atención	2090	5.84%
Lista de espera	1970	5.50%
Trato personal	906	2.53%
Hostelería, habitabilidad, confort	716	2%
Deficiencia información	697	1.94%
Deficiencia documentación	684	1.91%
Total	35772	100%

En la Figura.4 mostramos la evolución temporal de las quejas por cada motivo. Se observa que la evolución temporal de los motivos Lista de Espera y Trato Personal son bastante uniforme. Los motivos de *Demora en la lista de espera*, *Deficiencia en la Documentación* y *Desacuerdo en el Proceso Asistencial* si presentan un cambio temporal en su número de quejas ya que entre los años 2007 y 2008 alcanzan su máximo y en los años 2009 y 2010 desciende notablemente el número de quejas. Los motivos Hostelería, Habitabilidad y Confort, Deficiencia en la Información y Demora en la atención parecen oscilar en torno a un número medio de quejas.

Figura.4 Evolución anual de las quejas para cada motivo

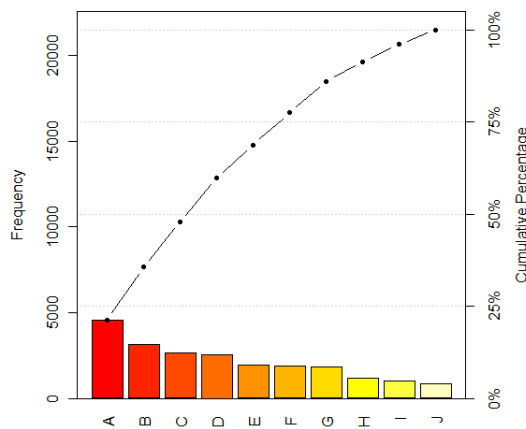


3.4. Servicios clínicos

Se recogen quejas en prácticamente todos los servicios con los que cuenta el Hospital. A continuación se muestran los resultados de los 10 servicios con mayor número de quejas y su distribución temporal.

En la Figura.5 presentamos el Diagrama de Pareto de los 10 servicios más relevantes, donde se puede observar el número total de quejas en dichos servicios y su porcentaje sobre el total de quejas.

Figura.5 Diagrama de Pareto por servicios clínicos



El servicio con mayor número de quejas es *el servicio A*, seguido de otros servicios como *servicio B*, *servicio C* y *servicio D*. *El servicio A* tiene casi un 13% de quejas sobre el total del hospital y entre los 10 servicios analizados aglutinan un 60% de las quejas. En la Tabla.5 mostramos todos los datos brutos.

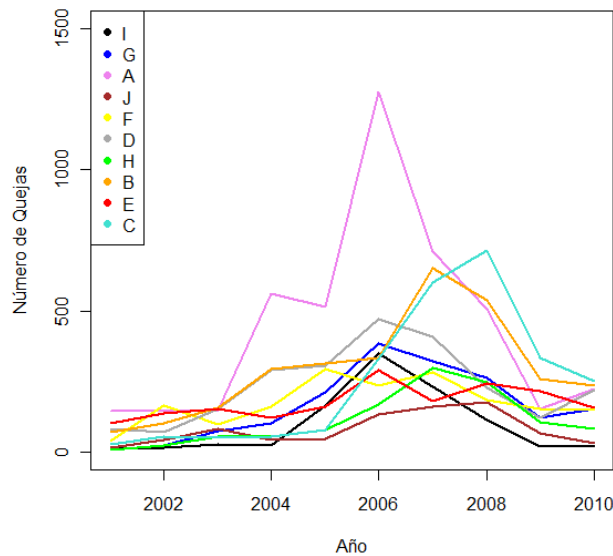
Tabla.5 Distribución de las quejas por servicios

Servicios	Número de quejas	Porcentaje
A	4545	12.71%
B	3121	8.72%
C	2628	7.35%
D	2558	7.15%
E	1926	5.38%
F	1886	5.27%
G	1815	5.07%
H	1167	3.26%
I	994	2.78%
J	845	2.36%
Total 10 servicios	21485	60.06%
Total	35772	100%

Como en el caso de los motivos de las quejas, se mostrará la evolución temporal de las quejas en cada servicio.

En la Figura.6 se muestra una comparativa de los servicios con mayor número de quejas. Se puede observar como casi todos estos servicios tienen el pico de quejas en el año 2006. Los servicios de *B* y *C*, que son los servicios con mayor número de quejas después del *servicio A*, tienen en los años 2007 y 2008 sus picos de quejas respectivamente.

Figura.6 Evolución anual de las quejas por servicios



Ya que el número de quejas depende del número de Consultas realizadas en cada servicio, calculamos una tasa del número de quejas por cada 10000 consultas.

En la Tabla.6 se muestran el número de quejas de cada servicio, su número de consultas y la tasa del número de quejas por 10000 consultas. En la Figura.7 se presenta el Diagrama de Pareto y en la Figura.8 la evolución temporal de las tasas.

Para este análisis tuvimos que prescindir de los servicios D, H e I ya que no tenemos la frecuentación de consultas. Como tampoco se disponía del número de Consultas del año 2011, el análisis de las quejas se excluyó este año.

Con este análisis de tasas, se observa que *el servicio C*, pasa a tener la tasa más elevada donde en números brutos ocupaba el tercer lugar mientras que el servicio A, que anteriormente era el que más quejas tenía ahora pasa a ser el servicio con la segunda mayor tasa de quejas.

Tabla.6 Distribución de las quejas y de las tasas por servicios

Servicios	Número de quejas	Consultas	Tasa (*10 ⁻⁴)
G	1673	316599	53
A	4385	713736	61
J	816	198555	41
F	1765	664231	27
B	2963	559257	53
E	1776	871121	20
C	2498	226447	110
Total	15876	3549946	365

Figura.7 Diagrama de Pareto de las tasas de los servicios

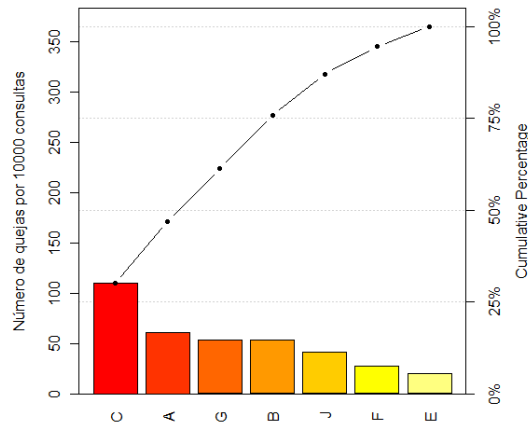
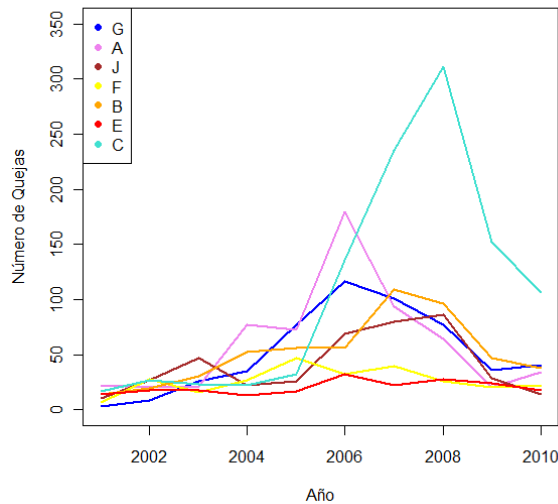


Figura.8 Evolución anual de las tasas de los servicios



3.5. Distribución de los diferentes motivos según servicios

En la Tabla.7a se muestran las quejas totales de cada servicio clínico desglosadas por motivos. La tabla 7.b se presentan los resultados en porcentajes de motivos de las quejas en cada servicio. Se observa que el motivo más frecuente en todo los servicios es *Demora en la lista de espera* menos en el servicio E en el que solo significa casi un 4% de las quejas, siendo *Demora en la atención* y *Descuerdo en el proceso asistencial* los más frecuentes.

Tabla.7a Distribución de las quejas de cada servicio por motivos

	I	G	A	J	F	D	H	B	E	C
Def. Información	2	5	32	7	19	32	2	29	80	7
Def. Doc.	9	9	34	11	32	73	9	91	18	9
Lista Espera	22	2	16	4	23	NA	1	58	NA	13
Dem. List. Espera	886	1677	4101	739	1467	2200	1087	2503	75	2503
Demora Atención	28	33	56	14	63	97	3	119	918	23
Desac. Proc.Asist.	36	71	238	55	236	98	60	261	539	69
Trato Personal	8	17	60	15	34	41	5	56	181	3
Host. Habit. Conf	3	1	8	0	12	17	0	4	115	1
Total	994	1815	4545	845	1886	2558	1167	3121	1926	2628

Tabla.7b Distribución del porcentaje de los motivos de las quejas en cada servicio

	I	G	A	J	F	D	H	B	E	C
Def. Información	0.2	0.28	0.7	0.83	1.01	1.26	0.17	0.93	4.15	0.27
Def. Doc.	0.92	0.5	0.75	1.3	1.7	2.85	0.77	2.92	0.94	0.34
Lista Espera	2.21	0.11	0.35	0.47	1.22	0	0.09	1.86	0	0.49
Dem. List. Espera	89.13	92.39	90.23	87.45	77.78	86	93.14	80.2	3.89	95.24
Demora Atención	2.82	1.82	1.23	1.66	3.34	3.79	0.26	3.81	47.66	0.88
Desac. Proc.Asist.	3.62	3.9	5.24	6.51	12.51	3.83	5.14	8.36	27.99	2.63
Trato Personal	0.8	0.94	1.32	1.78	1.8	1.6	0.43	1.79	9.4	0.11
Host. Habit. Conf	0.3	0.06	0.18	0	0.64	0.67	0	0.13	5.97	0.04
Total	100	100	100	100	100	100	100	100	100	100

NA: No Aplicable.

4. CONCLUSIONES

Los años 2006 y 2007 son los años con mayor número de quejas. A partir del 2007 el número de quejas comienza a descender y en los años 2009 y 2010 vuelven a situarse al nivel del 2004-2005.

Los meses de primavera y otoño, son en los que se producen un mayor número de quejas, mientras que los meses de invierno y verano disminuyen las quejas.

Dos terceras partes de las quejas se producen en el grupo de quejas *Demora en la lista de espera*.

Entre los 10 servicios donde más quejas se producen, estos aglutinan un 60% del total de las quejas del hospital, el servicio A es el que más tiene.

Ya que el número de quejas depende del número de consultas realizadas en cada servicio, calculamos una tasa del número de quejas por cada 10000 consultas. Con este análisis de tasas, se observa que *el servicio C*, pasa a tener la tasa más elevada donde en números brutos ocupaba el tercer lugar mientras que el servicio A, que anteriormente era el que más quejas tenía ahora pasa a ser el servicio con la segunda mayor tasa de quejas.

En todos los servicios, menos en el *servicio C*, la queja que predomina es *Demora lista de espera*. En el *servicio C* los motivos de quejas más frecuentes son: *Demora en la atención y Desacuerdo en el proceso asistencial*.

Capítulo 2

CONCENTRACIÓN DE ELASTASA LEUCOCITARIA EN LA ENFERMEDAD ARTERIAL CORONARIA DIAGNOSTICADA MEDIANTE ANGIOGRAFÍA

1. INTRODUCCIÓN

Aunque existen numerosas técnicas no invasivas capaces de detectar enfermedad arterial coronaria, (coronary artery disease CAD), como por ejemplo, electrocardiograma (con o sin esfuerzo), ecocardiograma (con o sin estrés) y el uso de radiotrazadores. Todos ellos comparten una característica, detectan síntomas derivados de la isquemia más que las placas de ateroma, que son la causa de esa isquemia. Una consecuencia es que estos métodos tienen un éxito muy limitado en la detección de CAD. Una posible solución puede ser la de medir factores bioquímicos relacionados con la formación o destrucción de la placa de ateromas.

La elastasa es una enzima proteolítica liberada por los neutrófilos, que se ha visto implicada en la fisiopatología de la enfermedad isquémica del corazón.

El objetivo de este estudio es investigar si la elastasa leucocitaria en sangre puede asociarse a la enfermedad arterial coronaria, y si los niveles de esta enzima pueden ser útiles en su diagnóstico.

2. MÉTODOS

Sujetos

En este estudio se incluyeron 141 pacientes, reclutados en el Servicio de Cardiología del Complejo Hospitalario Universitario de Santiago de Compostela (CHUS), y que se sometieron a una angiografía coronaria para la evaluación de dolor precordial o de una enfermedad cardíaca valvular. Ninguno de los sujetos estaba tomando antibióticos y no tenían historia clínica que pudiesen sugerir neumopatías, neoplasia, enfermedad cerebrovascular, enfermedad arterial periférica o enfermedad sistémica.

De los 141 pacientes, 45 tuvieron un angiograma coronario normal (grupo no-CAD) y 96 tuvieron un angiograma coronario anormal (grupo CAD). En este último grupo, a su vez, las lesiones coronarias fueron clasificadas como simples y como complejas.

Variables

Las variables de nuestro estudio son: la edad, el sexo, la talla, el peso, la concentración total de colesterol, la concentración de elastasa en los pacientes y si estos tienen o no hipertensión, diabetes, angina, si son fumadores o no y si tienen un angiograma coronario anormal o normal (CAD o no-CAD).

Análisis estadístico

La asociación entre cada una de las variables del estudio y el status coronario (no-CAD o CAD) fueron evaluadas usando el t-test y el test de Wilcoxon para variables continuas, y para comprobar la independencia de las variables categóricas utilizamos el test exacto de Fisher y el test chi-cuadrado.

El efecto de posibles variables de confusión se evaluó utilizando un modelo logístico de regresión múltiple, usando un procedimiento stepwise. Solamente las variables significativas fueron guardadas en el modelo. La estimación de los riesgos relativos y de los intervalos de confianza al 95% fueron calculados a partir de los coeficientes.

Para el cálculo de los diferentes indicadores que permiten el conocimiento de la capacidad diagnóstica de un test, se utilizó la librería DiagnosisMed.

3. RESULTADOS

3.1. Características clínicas de los pacientes del estudio

La Tabla.1 muestra las características clínicas del grupo CAD y del grupo no-CAD (grupo control). El grupo no-CAD es más joven que el grupo CAD ($P < 0.05$) contiene menos hombres (alrededor del 51% en contra del 84% en el grupo CAD; $P < 0.001$) y un menor número de fumadores y enfermos de angina de pecho.

El grupo control es menos alto, menos pesado y tiene una menor concentración de colesterol que el grupo CAD y contiene un número menor de hipertensos y diabéticos, aunque estas diferencias no fueron significativas.

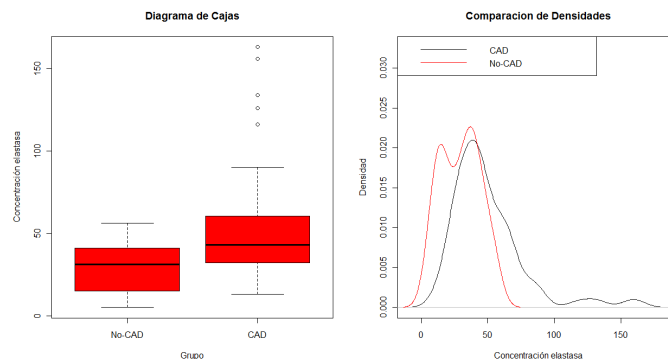
La concentración de elastasa del grupo CAD ($49.7 \pm 2.8 \mu\text{g.l}^{-1}$) y del grupo No-CAD ($29.5 \pm 2.2 \mu\text{g.l}^{-1}$) difieren significativamente ($P < 0.001$), podemos verlo en la Figura.1.

Tabla.1 Características clínicas de los pacientes

	Grupo No-CAD (n=45)	Grupo CAD (n=96)	P-valor
Edad (años)	56.3 ± 1.8	62.3 ± 1.0	0.0064
Sexo (Hombre)	23 (51%)	81 (84%)	0.0000
Talla (cm)	162.2 ± 1.2	163 ± 0.7	0.5938
Peso (kg)	73.5 ± 1.8	73.8 ± 1.2	0.8437
Hipertensión (Si)	12 (29%)	28 (31%)	0.8951
Diabetes (Si)	2 (5%)	15 (17%)	0.0915
Fumador (Si)	16 (38%)	55 (62%)	0.0153
Angina (Si)	24 (53%)	94 (98%)	0.0000
Colesterol	207.7 ± 6.3	210.2 ± 4.7	0.7481
Elastasa	29.5 ± 2.2	49.7 ± 2.8	0.0000

En cada grupo para las variables cuantitativas se representa la media y su error estándar y para las variables cualitativas las frecuencias y sus porcentajes.

Figura.1 Diagrama de cajas y las diferentes densidades de elastasa en ambos grupos



3.2. Características clínicas del grupo CAD

La Tabla.2 muestra las características clínicas del grupo CAD, diferenciando los tipos de lesiones coronarias (Simples y Complejas).

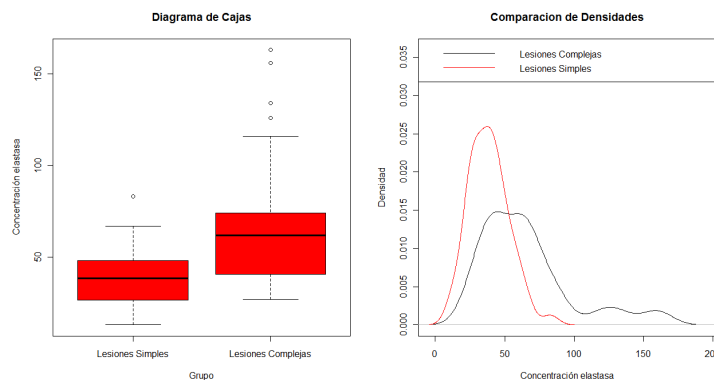
La concentración de elastasa en lesiones simples ($38.6 \pm 1.9 \mu\text{g.l}^{-1}$) difiere significativamente ($P < 0.001$), de la concentración de elastasa en lesiones complejas ($65.2 \pm 5.3 \mu\text{g.l}^{-1}$), podemos verlo en la Figura.2.

Tabla.2 Características clínicas del grupo CAD

	Lesiones Simples (n=56)	Lesiones Complejas (n=40)	P-valor
Edad (años)	61.0 ± 1.3	64.1 ± 1.7	0.1331
Sexo (Hombre)	47 (84%)	34 (85%)	0.8867
Talla (cm)	163.7 ± 1.0	161.8 ± 1.0	0.1110
Peso (kg)	75.1 ± 1.5	71.8 ± 1.9	0.1613
Hipertensión (Si)	14 (27%)	14 (37%)	0.4759
Diabetes (Si)	6 (12%)	9 (24%)	0.2303
Fumador (Si)	35 (69%)	20 (54%)	0.2416
Colesterol	215.3 ± 6.1	201.7 ± 7.0	0.1482
Elastasa	38.6 ± 1.9	65.2 ± 5.3	0.0000

En cada grupo para las variables cuantitativas se representa la media y su error estándar y para las variables cualitativas las frecuencias y sus porcentajes.

Figura.2 Diagrama de cajas y las densidades de elastasa en ambas lesiones



3.3. Análisis de la asociación entre la concentración de la elastasa y la enfermedad arterial coronocaria (CAD)

Para evaluar la asociación entre la elastasa y la enfermedad coronaria, ajustando por el efecto de posibles variables de confusión, se utilizó un modelo de regresión logística, usando el procedimiento “stepwise”.

Las variables que tienen un valor significativo independiente para la predicción del CAD entre pacientes incluidos en el estudio fueron: angina, sexo, edad y la concentración de elastasa. En resumen partimos de un modelo con las 10 variables iniciales que aparecen en la Tabla.1, y eligiendo un modelo, utilizando un algoritmo paso a paso por AIC, nos quedamos con las variables: angina, sexo, edad y la concentración de elastasa.

En la Tabla.3 se muestra los coeficientes estimados del modelo de regresión logística, el error estándar, el p-valor (si es menor que 0.05, indica que ese coeficiente es significativo en el modelo), el riesgo relativo y su correspondiente intervalo de confianza al 95%.

Tabla.3 Estimación de los coeficientes del modelo

	Coeficientes	SE	P-valor	Riesgo Relativo (R.R)	I.C. 95% del R.R
Constante	-11.29	2.77	0.0000		
Angina	5.24	1.31	0.0001	188.14	
Sexo	2.15	0.72	0.0029	8.57	(2.26, 40.33)
Edad	0.04	0.03	0.1741	1.04	(0.98, 1.10)
Elastasa	0.10	0.02	0.0000	1.11	(1.06, 1.17)

Viendo los resultados de la Tabla.3 podemos decir que los que tienen angina, tienen 188 veces más riesgo de tener CAD que los que no presentan una angina de pecho. Los hombres tienen 8.57 veces más riesgo de tener CAD que las mujeres. Y que el riesgo de CAD se incrementa un 11% por cada unidad de aumento en la elastasa.

3.4. Capacidad diagnóstica de un test

Anteriormente hemos visto que la elastasa está asociada a la enfermedad arterial coronaria. El segundo objetivo que deseamos contestar es el de evaluar la capacidad diagnóstica de la elastasa en la enfermedad arterial coronaria.

Habitualmente para estas evaluaciones se utiliza una discriminación que tenga la capacidad de separar aquellos individuos enfermos o sanos. Antes de introducir la medida de discriminación más utilizada definiremos dos conceptos previos, que son los siguientes:

- **Sensibilidad (*sen*):** Indica la capacidad de la prueba para detectar un sujeto enfermo.
- **Especificidad (*esp*):** Indica la capacidad de identificar como sanos a los individuos que realmente los son.

La confianza que realiza el test será mayor cuando la sensibilidad y especificidad del test estén más próximas a 1, siendo el test ideal aquel con $sen=1$ y $esp=1$, sin embargo en la práctica esta situación ideal nunca se produce.

La medida de discriminación más utilizada es la curva *ROC* (Receiver Operating Characteristic), que se define (Swets y Pickett, 1982) como el conjunto de posibles valores de $1-esp(c)$ (falsos positivos) y $sen(c)$ (verdaderos positivos) que en nuestro caso, se alcanzan para cada valor de la concentración de elastasa.

La precisión de un test será mayor cuanto más próxima este la curva *ROC* a la esquina superior izquierda del cuadrado. Esto sugiere la utilización del área bajo la curva *ROC* (*AUC*, Area under the curve) como medida de la capacidad diagnóstica de un test.

Además de la sensibilidad y la especificidad existen otros indicadores que permiten el conocimiento casi completo de la capacidad diagnóstica del test, y son los siguientes:

- **Índice de Validez (*IV*):** Se define como la proporción de individuos clasificados correctamente, la utilización de este índice tiene un problema que es que depende de la prevalencia de la enfermedad.
- **Índice de Youden (*IY*):** Refleja la diferencia entre la tasa de verdaderos positivos y la de falsos positivos, por lo tanto cuanto más cercano a 1 sea el *IY*, mejor es la prueba diagnóstica que se está evaluando.
- **Razón de Verosimilitud (*RV*):** Es un índice que cuantifica cuantas veces es más probable que el test sea positivo en los enfermos que en los sanos. La *RV* viene dado por la siguiente expresión matemática $RV+ = \frac{sen}{1-esp}$ cuyo complementario es $RV- = \frac{1-sen}{esp}$ razón de verosimilitud para resultados negativos. Un buen test debe tener una *RV* - cercana a 0 y una *RV* + alta (no es posible especificar un límite superior).

Para calcular: los puntos de corte de la variable elastasa (*PC*), los verdaderos positivos (*VP*), los falsos negativos (*FN*), los falsos positivos (*FP*), los verdaderos negativos (*VN*), la sensibilidad (*sen*), la especificidad (*esp*), el valor predictivo positivo (*VPP*), el valor predictivo negativo (*VPN*), la razón de verosimilitud para resultados positivos (*RV+*), la razón de verosimilitud para resultados negativos (*RV-*), el Índice de Youden (*IY*) y los intervalos de confianza (IC) del 95% para *sen*, *esp*, *VPP*, *VPN*, *RV+* y *RV-* utilizamos la librería *DiagnosisMed*.

Usaremos tres criterios diferentes:

- “*Se=Sp*” (1) es el criterio en el que para un determinado valor de la elastasa, se iguala la sensibilidad y la especificidad.
- “*Max.Youden*” (2) es el criterio en el cual un determinado valor de la variable elastasa, maximiza el Índice de Youden.
- “*Max.DOR*” (3) es el criterio en el que un determinado valor de la elastasa, maximiza la razón de probabilidad del diagnóstico.

Esos valores de la elastasa serán los puntos de corte para cada criterio.

En la Figura.3, Figura.4 y Figura.5 mostramos la curva *ROC* y la *AUC*, el punto de corte, la sensibilidad y la especificidad para los criterios (1), (2) y (3) respectivamente.

En la Figura.6, Figura.7 y Figura.8 representaremos la sensibilidad y la especificidad para cada valor de la concentración de elastasa y dibujaremos el punto de corte para los criterios (1), (2) y (3) respectivamente.

Finalmente en la Tabla.4 mostramos para cada punto de corte calculado por los criterios (1), (2) y (3): *VP*, *FN*, *FP*, *VN* e *IY* y en la Tabla.5 mostramos para cada punto de corte calculado por los criterios (1), (2) y (3): *sen*, *esp*, *VPP*, *VPN*, *RV+* y *RV-* y sus correspondientes intervalos de confianza (IC) al 95%.

Figura.3 Curva ROC y punto de corte calculado por el criterio $Se=Sp$

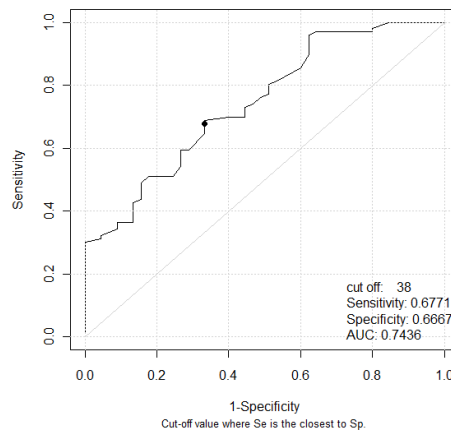


Figura.4 Curva ROC y punto de corte calculado por el criterio Max. Youden

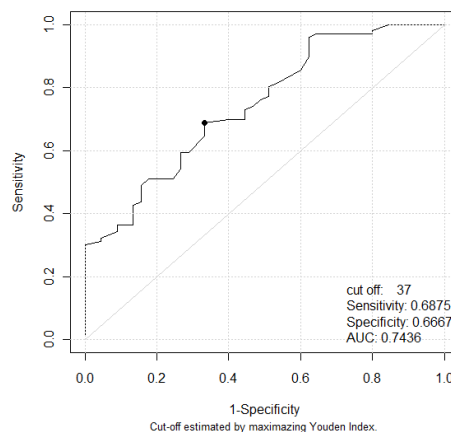


Figura.5 Curva ROC y punto de corte calculado por el criterio Max.DOR

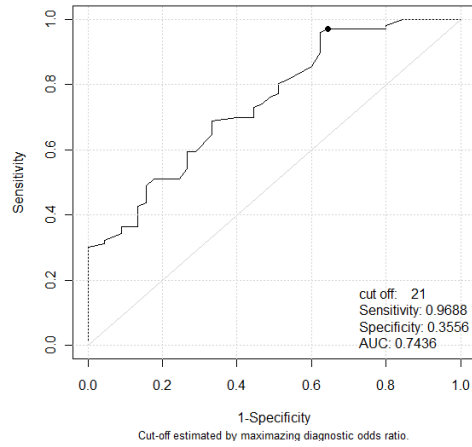


Figura.6 Gráfico de la sensibilidad, especificidad y punto de corte obtenidos por el criterio $Se=Sp$

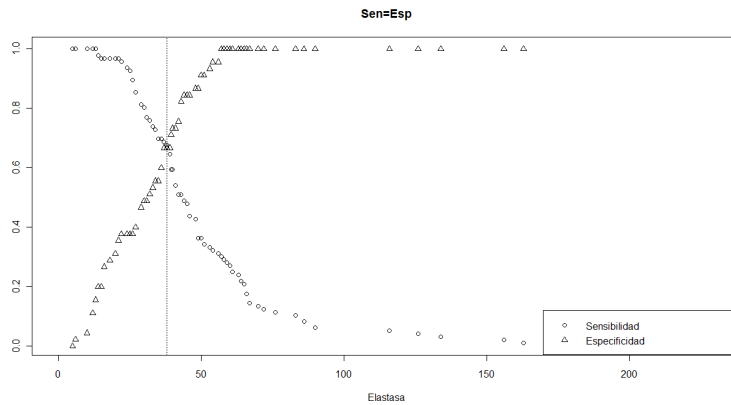


Figura.7 Gráfico de la sensibilidad, especificidad y punto de corte obtenidos por el criterio Max.Youden

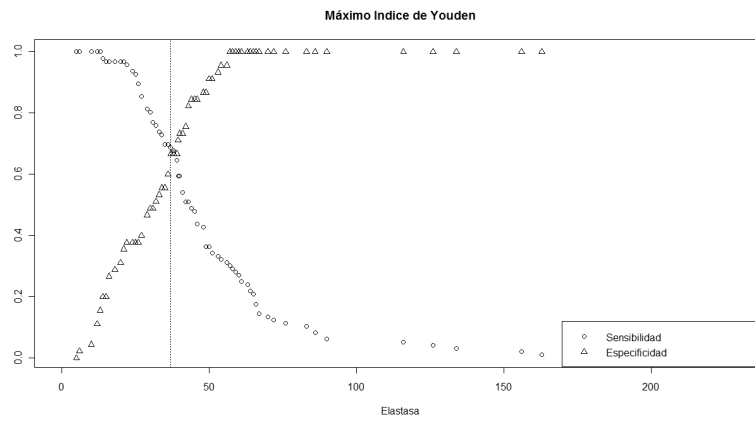


Figura.8 Gráfico de la sensibilidad, especificidad y punto de corte obtenidos por el criterio Max.DOR

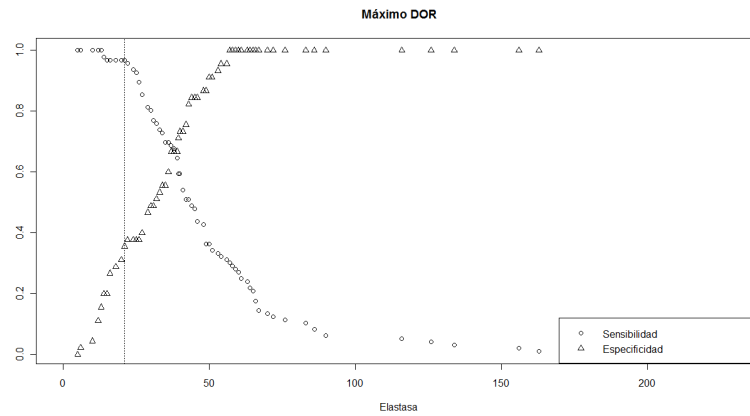


Tabla.4 Indicadores de la capacidad diagnóstica para cada uno de los criterios

	PC	VP	FN	FP	VN	VPP	VPN	IY
“Se=Sp” (1)	38	65	31	15	30	0.8125	0.4918	0.3438
“Max.Youden” (2)	37	66	30	15	30	0.8148	0.5000	0.3542
“Max.DOR” (3)	21	93	3	29	16	0.7623	0.8421	0.3244

Tabla.5 Indicadores de la capacidad diagnóstica, y sus IC al 95% para cada uno de los criterios

	PC	Sen	Esp	RV+	RV-
“Se=Sp” (1)	38	0.68 (0.58, 0.76)	0.67 (0.52, 0.79)	2.03 (1.31, 3.14)	0.48 (0.34, 0.69)
“Max.Youden” (2)	37	0.69 (0.59, 0.77)	0.67 (0.52, 0.79)	2.06 (1.33, 3.18)	0.47 (0.33, 0.67)
“Max.DOR” (3)	21	0.97 (0.91, 0.99)	0.36 (0.23, 0.5)	1.5 (1.2, 1.88)	0.09 (0.03, 0.29)

La estimación no paramétrica del área es 0.7436, cuyo IC al 95% es (0.6772, 0.8181), por lo tanto podemos decir que la elastasa discrimina bastante bien a individuos enfermos (CAD) y sanos (No-CAD).

En nuestro caso, el mejor criterio sería el (3) ya que clasifica a 3 individuos que tienen CAD como si no lo tuviesen (Falsos Negativos, *FN*) y clasifica a 29 individuos que no tienen CAD como si lo tuviesen (Falsos Positivos, *FP*), mientras que el criterio (1) y (2) clasifican respectivamente a 31 y 30 individuos que tienen CAD como si no lo tuviesen y clasifican a 15 individuos que no tienen CAD como si lo tuviesen.

El criterio (3) comete menos falsos negativos que los criterios (1) y (2) por lo que en caso de error, para nuestro caso, es menos grave que siga hospitalizado un sano que dar el alta a un enfermo.

La probabilidad de un resultado positivo para el criterio (3) es 1.5 veces mayor en los pacientes con CAD que en los pacientes que no tienen CAD y en cambio la probabilidad de un resultado negativo es 11 veces superior ($1/0.09=11.1$) en los pacientes con no-CAD que en los que tienen CAD. Este test tiene una sensibilidad alta pero una especificidad muy baja lo que confiere mayor valor para detectar enfermos que para descartar sanos.

4. CONCLUSIONES

Las características clínicas de los pacientes del estudio son las siguientes:

- La edad media en el grupo CAD es significativamente mayor que la del grupo No-CAD.
- La proporción de: hombres, personas diabéticas y personas que tienen una angina de pecho en el grupo CAD, es significativamente mayor que en el grupo No-CAD.
- La concentración de elastasa media en el grupo CAD es significativamente mayor que la del grupo No-CAD.

La característica clínica dentro del grupo CAD, diferenciando los dos tipos de lesiones coronarias (Simples y Complejas), es la siguiente:

- La concentración de elastasa media en el grupo de lesiones complejas es significativamente mayor que la del grupo de lesiones simples.

Podemos decir que el riesgo de CAD se incrementa un 11% por cada unidad de aumento de la concentración de elastasa.

La estimación no paramétrica del *AUC* es 0.7436, cuyo IC al 95% es (0.6772, 0.8181), por lo tanto la elastasa discrimina bastante bien a individuos enfermos (CAD) y sanos (No-CAD).

Eligiendo, mediante el criterio que maximiza la razón de probabilidad del diagnóstico ("Max.DOR"), un punto de corte ($PC=21$) para la concentración de elastasa, obtenemos un test que tiene una sensibilidad alta pero una especificidad muy baja lo que confiere mayor valor para detectar enfermos que para descartar sanos.

Capítulo 3

FUNCIONES EN R PARA UN ANÁLISIS DESCRIPTIVO Y COMPARATIVO DE VARIABLES CUALITATIVAS Y CUANTITATIVAS

1. INTRODUCCIÓN

En la Unidad de Epidemiología del Complejo Hospitalario de Santiago de Compostela (CHUS) es frecuente el uso de variables cualitativas (Sexo, Fumador, Anemia, Diabetes...) y cuantitativas (Edad, Peso, Colesterol...) de ahí la utilidad de crear unas funciones en R, que nos faciliten el trabajo de “describir” esos datos, ya que es el primer paso en cualquier análisis que se realice.

En general, y en concreto en esta unidad de epidemiología es frecuente estudiar si un conjunto de variables continuas (independientes o relacionadas) presenta distribuciones iguales. También es muy frecuente comprobar si dos variables cualitativas son independientes y contrastar si dos variables cualitativas relacionadas son simétricas (por ejemplo, si existen cambios significativos de la variable de interés entre dos tiempos, basal y final).

Las variables según su naturaleza se puede clasificar en:

- **Cualitativas:** No se pueden medir numéricamente, por ejemplo sexo o nivel de estudios. Cuando las categorías de la variable (Hombre/Mujer; Sin estudios/Estudios primarios/Estudios secundarios) no tiene orden, la variable se dice que es nominal (sexo), y si se puede ordenar se dice que es ordinal (nivel de estudios).
- **Cuantitativa:** Cuando puede tomar valores numéricos. El número de hijos o la estatura de un individuo son variables cuantitativas. Se clasifican en discretas y continuas. Una variable se dice que es discreta si solo puede tomar un número finito de valores o infinito numerable (número de hijos). Por el contrario, las variables continuas pueden tomar todos los valores de un determinado intervalo (la estatura).

Es muy importante saber distinguir entre los distintos tipos de variables, ya que los análisis entre variables cualitativas y cuantitativas se realizan de manera completamente diferente.

2. FUNCIONES EN R PARA UN ANÁLISIS DESCRIPTIVO DE LOS DATOS

Las funciones *des.cont* y *des.cat.def* realizan un análisis descriptivo de una variable cuantitativa y cualitativa respectivamente. El único argumento que hay que pasarle a las funciones es la variable que se desea “describir”.

Al ejecutar la función *des.cont* nos aparece en pantalla:

- El número de elementos de la variable que se está describiendo.
- La cantidad de datos perdidos (NA's) que tiene esa variable.
- La media y entre paréntesis la desviación típica (standard deviation, sd) y el error estándar (standard error, se) que se calcula de la siguiente manera: $\hat{\sigma} / \sqrt{n}$.
- La mediana y entre paréntesis el primer cuartil (Q_1), el tercer cuartil (Q_3) y el rango intercuartílico (Interquartile Range, IQR) que se calcula de la siguiente manera: $Q_3 - Q_1$.

Al ejecutar la función *des.cat.def* nos aparece en pantalla:

- La cantidad de datos perdidos que tiene esa variable cualitativa.
- Para cada categoría de la variable cualitativa nos aparece sus frecuencias y al lado de ellas entre paréntesis sus porcentajes.

3. FUNCIONES EN R PARA UN ANÁLISIS COMPARATIVO DE LOS DATOS

Una vez que hayamos identificado la naturaleza de las variables del estudio construiremos unas funciones que nos permitan comparar variables cualitativas y variables cuantitativas. La función *test.cualitativa* compara dos variables cualitativas, la función *test.cuantitativa* compara dos variables cuantitativas y la función *test.mixto* analiza una variable cuantitativa y una cualitativa.

Pero antes explicaremos el funcionamiento de dos funciones *normalidad* y *tablas.bidim*. La primera función *normalidad* se utiliza en *test.mixto* y en *test.cuantitativa* y lo que hace es darnos el p-valor del Test de Lilliefors (Lilliefors, 1967) si el tamaño de la muestra es mayor o igual que 50. Si el tamaño es menor que 50, nos devuelve el p-valor del Test de Shapiro-Wilk, (Shapiro y Wilk, 1965). El único argumento que hay que pasarle a esta función es la variable cuantitativa, que se desea contrastar.

La función *tablas.bidim* se utiliza en *test.cualitativa* y al ejecutarla nos aparece en pantalla:

- La tabla de frecuencias absolutas y la tabla de frecuencias relativas correspondiente a dos variables cualitativas.
- Las distribuciones condicionales por columnas (si las dos variables cualitativas no pertenecen al mismo individuo) o por filas (si las dos variables categóricas pertenecen al mismo individuo, es decir si son relacionadas).

Los argumentos que hay que pasarle a esta función son: el nombre de las dos variables cualitativas, los datos, y si están o no relacionadas esas dos variables (*paired=TRUE* o *FALSE*).

3.1. Función en R para un análisis comparativo de dos variables cualitativas

Uno de los principales objetivos de la función *test.cualitativa* será contrastar la hipótesis de independencia entre dos variables cualitativas basándonos en la información proporcionada por las frecuencias observadas contenidas en la tabla de contingencia.

Dentro de esta función se ejecuta la función *des.cat.def* para “describir” las dos variables cualitativas y luego la función *tablas.bidim* para calcular las tablas de frecuencias absolutas y relativas y las distribuciones condicionales por columnas o filas.

Para tablas 2x2 tenemos dos opciones utilizar el contraste de independencia chi-cuadrado o el test exacto de Fisher.

Para poder aplicar correctamente el contraste de independencia chi-cuadrado, debe cumplirse que la muestra sea grande ($n > 30$) y las frecuencias esperadas mayores o iguales que 5.

En el caso de un tamaño de muestra pequeño, debe utilizarse la corrección por continuidad propuesta por Yates (Yates, 1934) para el estadístico chi-cuadrado. Esta corrección solo es válida para el caso de tablas 2x2. Existe mucha polémica sobre la aplicación de la corrección de Yates porque hay casos prácticos en los que con el estadístico corregido se rechaza la independencia con bastante menor significación que con el test no corregido. Es decir, el p-valor corregido es bastante mayor que el no corregido, por lo que el test corregido es más conservativo. Cabe destacar que en caso de tamaño muestral grande, la corrección tiene efecto pequeño sobre el valor del estadístico, por ello, no utilizaremos dicha corrección en este caso.

En esta función se calcula los estimadores de máxima verosimilitud de las frecuencias esperadas bajo la hipótesis nula de independencia. Se calculan de la siguiente manera: $\hat{m}_{ij} = n_i n_j / n$, donde n_i y n_j son las sumas de la *i*-ésima fila y de la *j*-ésima columna respectivamente, de la tabla de frecuencias absolutas y n es el tamaño de la muestra.

En el caso de que todas las frecuencias esperadas sean mayores o iguales que 5 utilizamos el contraste de independencia chi-cuadrado (con o sin corrección dependiendo del tamaño muestral) y si alguna de estas frecuencias son menores que 5 utilizamos el test exacto de Fisher (Fisher, 1934).

Para el caso general que tengamos tablas $I \times J$, quitando el caso en el que $I = 2$ y $J = 2$, utilizamos el contraste de independencia chi-cuadrado (con o sin corrección dependiendo del tamaño muestral).

Si los datos categóricos proceden de muestras relacionadas (por ejemplo medidas repetidas de un mismo individuo), las tablas que resultan de la representación de estos datos son cuadradas y además con las mismas categorías para filas y columnas. Para ellas el estudio de la independencia no tiene interés por lo que otro de los objetivos de la función *test.cualitativa* será contrastar la simetría (si existen cambios significativos de la variable de interés en el tiempo, por ejemplo basalmente y finalmente) y la homogeneidad marginal (si las diferencias entre las proporciones marginales son significativas).

Para tablas 2x2 ambos conceptos (simetría y homogeneidad marginal) son equivalentes. Para este tipo de tablas utilizaremos el test de McNemar (McNemar, 1947). Los estimadores de máxima verosimilitud de las frecuencias esperadas bajo simetría se calculan de la siguiente manera:

$$\hat{m}_{11} = n_{11}; \hat{m}_{12} = n_{12}; \hat{m}_{21} = \hat{m}_{22} = (n_{12} + n_{21})/2.$$

Al igual que en el test chi-cuadrado, la corrección de Yates es aconsejable cuando las frecuencias esperadas en el caso de simetría son menores que 5, es decir, $(n_{12} + n_{21})/2 < 5$.

Consideramos ahora tablas cuadradas $I \times I$, es decir cuando la variable cualitativa tiene tres o más categorías. En este caso la simetría y la homogeneidad marginal no son equivalentes (para tablas 2x2 sí que lo son), sin embargo la simetría implica homogeneidad marginal pero el recíproco no es cierto.

Si queremos contrastar la simetría para este tipo de tablas utilizaremos la extensión de Bowker (Bowker, 1948) para al test de McNemar. Si la hipótesis de simetría es rechazada puede ser de interés contrastar la hipótesis de homogeneidad marginal por lo que utilizaremos la extensión de Stuart (Stuart, 1955) y Maxwell (Maxwell, 1970) al test de McNemar. Para realizar este test se debe instalar y cargar la librería coin.

Los argumentos que hay que pasarle a la función *test.cualitativa* son: los nombres de las dos variables cualitativas, los datos y si están relacionadas o no esas dos variables (paired=TRUE o FALSE). Además esta función representa un diagrama de barras para los datos.

3.2. Función en R para un análisis comparativo de dos variables cuantitativas

En la función *test.cuantitativa* se analizan dos variables cuantitativas dependiendo si proceden de muestras independientes o dependientes, también comprobaremos si las dos variables son normales para realizar test paramétricos o no paramétricos. El principal objetivo de esta función es contrastar si las dos variables tienen la misma distribución.

Dentro de esta función se ejecuta *des.cont* para “describir” las dos variables cuantitativas.

Para comprobar la normalidad de los datos, dentro de esta función ejecutamos la función *normalidad*, si la variable tiene más de 50 elementos realizamos el test de Lilliefors (recordemos que para realizar este test hay que cargar e instalar la librería *nortest*) y si tiene menos de 50 elementos realizamos el test de Shapiro-Wilk.

Por lo que si las variables son independientes y normales, en primer lugar se realiza la prueba F de comparación de varianzas. Si las dos variables tienen la misma varianza se realiza la prueba t de comparación de medias (con varianzas iguales) y si no tienen varianzas iguales se realiza la prueba t de comparación de medias (con varianzas diferentes).

Si las variables son independientes pero alguna de las dos no es normal realizamos la prueba no paramétrica de Wilcoxon.

Si las variables están relacionadas, es decir si por ejemplo son medidas repetidas en los individuos, y provienen de una distribución normal, realizaremos la prueba t de comparación de medias para muestras relacionadas, y si son variables relacionadas pero alguna de ellas no es normal realizamos la prueba no paramétrica de Wilcoxon para muestras relacionadas o dependientes.

Los argumentos que hay que pasarle a *test.cuantitativa* son: las dos variables continuas, los datos y si están o no relacionadas (*paired=TRUE* o *FALSE*). Además esta función representa un diagrama de cajas y la estimación tipo kernel de la función de densidad (función *density* en R) de cada una de las variables continuas.

3.3. Función en R para un análisis comparativo de una variable cuantitativa y otra cualitativa

En la función *test.mixto* analizaremos una variable cuantitativa y una variable cualitativa.

Dentro de esta función, en primer lugar se ejecuta *des.cont* y *des.cat.def* para “describir” una variable continua y categórica respectivamente.

Dependiendo de las k categorías que tenga la variable cualitativa tendremos k variables cuantitativas.

Si la variable es dicotómica, es decir que tiene $k=2$ categorías, el análisis es similar al que hace la función *test.cuantitativa*. Si la variable es policotómica, es decir que tiene $k>2$ categorías, tenemos dos situaciones:

1) Si las k variables cuantitativas son independientes diferenciamos tres situaciones:

- Si las k variables provienen de una distribución normal (esto lo sabemos ejecutando la función *normalidad*) y si tienen la misma variabilidad (para ello utilizaremos el Test de Levene (Levene, 1960), función *levneTest* del paquete *car*) se realiza un análisis de la varianza de un factor (ANOVA de un factor). En este caso al efectuar el ANOVA, si hay diferencias entre las medias de las k variables, se realiza el contraste a posteriori de Tukey para conocer donde se encuentran esas diferencias.
- Si las k variables provienen de una distribución normal pero no tienen la misma variabilidad, se realiza el test de Kruskal-Wallis para k muestras independientes.
- Si algunas de las k variables cuantitativas no es normal, se realiza el test de Kruskal-Wallis para k muestras.

2) Si las k variables cuantitativas están relacionadas se realiza el Test de Friedman para k muestras relacionadas o dependientes.

Los argumentos que hay que pasarle a la función *test.mixto* son: el nombre de una variable continua, el nombre de una variable categórica, los datos y si las k variables están o no relacionadas (*paired=TRUE* o *FALSE*). Además esta función representa los diagramas de cajas y las estimaciones no paramétricas tipo kernel de las densidades según las k variables.

4. EJEMPLO DEL USO DE LAS FUNCIONES

Utilizaremos los datos de elastasa (Capítulo N°2) para comprobar el funcionamiento de las funciones *test.cualitativa* y *test.mixto*.

En primer lugar queremos comprobar si la variable *edad* es diferente según la variable *grupo* (CAD, No-CAD). Como la *edad* es una variable cuantitativa y la variable *grupo* es cualitativa usamos la función *test.mixto*.

Creamos los script en R: “*descriptiva_cualitativa*”, “*descriptiva_continua*”, “*comparacion_cualitativas*”, “*comparacion_cuantitativas*”, “*comparacion_mixtas*”, “*normalidad*” y “*bidimensional*”. En estos scripts se encuentran las funciones: *des.cat.def*, *des.cont*, *test.cualitativa*, *test.cuantitativa*, *test.mixto*, *normalidad* y *tablas.bidim* respectivamente.

```

source("descriptiva_cualitativa.R")
source("descriptiva_continua.R")
source("comparacion_cualitativas.R")
source("comparacion_continuas.R")
source("comparacion_mixtas.R")
source("normalidad.R")
source("bidimensional.R")

elas$cad<-factor(elas$cad, levels=c(1,0), labels=c("CAD","No CAD"))

test.mixto("edad","cad",elas,paired=F)
*****
Descriptiva edad
*****
Número de elementos: 141
NAs: 0
Media (sd, se)
60.4 (11.148, 0.94)
Mediana (1st Qu, 3rd Qu, IQR)
61 (56, 68, 12)
*****
Descriptiva cad
*****
NAs: 0
      [,1]
CAD     96 (68.09%)
No-CAD 45 (31.91%)
*****
categórica-nivel: CAD
*****
Número de elementos: 96
NAs: 0
Media (sd, se)
62.3 (10.207, 1.04)
Mediana (1st Qu, 3rd Qu, IQR)
62 (56.75, 70, 13.25)
*****
TEST DE LILLIEFORS para nivel = CAD
p-valor: 0.386
*****
*****
categórica-nivel: No-CAD
*****
Número de elementos: 45
NAs: 0
Media (sd, se)
56.36 (12.078, 1.8)
Mediana (1st Qu, 3rd Qu, IQR)
59 (52, 63, 11)
*****
TEST DE SHAPIRO-WILK para nivel = No-CAD
p-valor: 0.0032
*****
*****
***** TEST NO PARAMÉTRICOS PARA DOS MUESTRAS *****
*****
***** Test de Wilcoxon *****

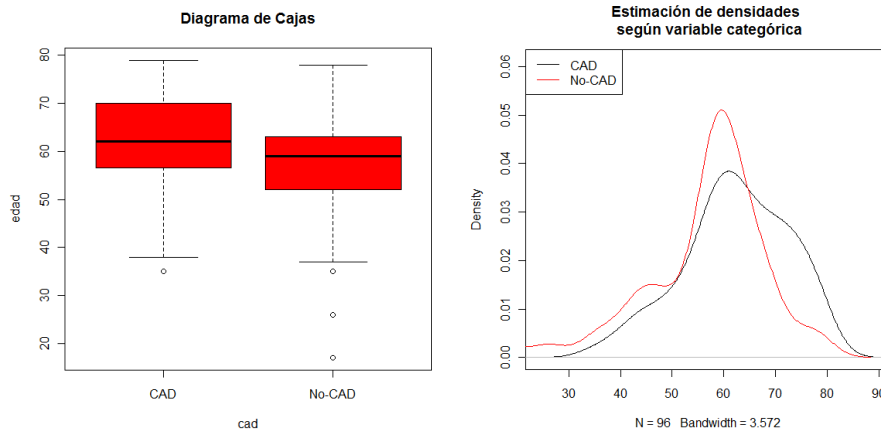
      Wilcoxon rank sum test with continuity correction

data: cont.var by cat.var
W = 2776, p-value = 0.006448
alternative hypothesis: true location shift is not equal to 0

NO TIENEN LA MISMA DISTRIBUCIÓN

```

Figura.1 Diagrama de cajas y las diferentes densidades de la edad en ambos grupos



En segundo lugar queremos comprobar si la variable *sexo* es independiente de la variable *grupo* (CAD, No-CAD). Como *sexo* y *grupo* son variables cualitativas usamos la función *test.cualitativa*.

```
elas$cad<-factor(elas$cad, levels=c(1,0), labels=c("CAD", "No CAD"))
elas$sexo<-factor(elas$sexo, levels=c(1,2), labels=c("Hombre", "Mujer"))
```

```
test.cualitativa("sexo", "cad", elas, paired=F)
```

```
*****
Descriptiva sexo
*****
NAS: 0
      [,1]
Hombre 104 (73.76%)
Mujer  37 (26.24%)
*****
Descriptiva cad
*****
NAS: 0
      [,1]
CAD    96 (68.09%)
No-CAD 45 (31.91%)
*****
Tabla de frecuencias absolutas
*****
      cad
sexo   CAD No-CAD
Hombre 81    23
Mujer  15    22
*****
Tabla de frecuencias relativas
*****
      cad
sexo   CAD    No-CAD
Hombre 0.5744681 0.1631206
Mujer  0.1063830 0.1560284
```

```

*****
Distribución sexo/cad = CAD
*****
NAs: 0
      [,1]
Hombre 81 (84.38%)
Mujer  15 (15.62%)
*****
Distribución sexo/cad = No-CAD
*****
NAs: 0
      [,1]
Hombre 23 (51.11%)
Mujer  22 (48.89%)
*****
***** TEST CHI-CUADRADO *****
*****

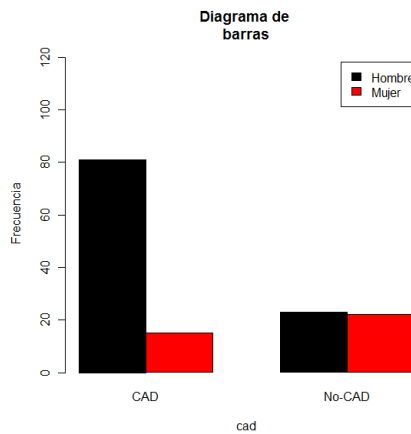
      Pearson's Chi-squared test

data:  cat1.var and cat2.var
X-squared = 17.5151, df = 1, p-value = 2.85e-05

LAS VARIABLES SON DEPENDIENTES

```

Figura.2 Diagrama de barras de las frecuencias absolutas del sexo en ambos grupos



Se puede comprobar que los resultados coinciden con la Tabla.1 del capítulo N°2, página N°21 de este informe.

Utilizaremos los datos de Insuficiencia Cardíaca (Capítulo N°4) para comprobar el funcionamiento de la función *test.cuantitativa*. Queremos comprobar si hay cambios en el nivel de hemoglobina basal (*hb0*) y hemoglobina final (*hb_fin*) en unos individuos seguidos en una consulta.

```

test.cuantitativa("hb0","hb_fin",Seguidos,paired=T)
*****
Descriptiva hb0
*****
Número de elementos: 96
NAs: 0
Media (sd, se)
12.91 (1.8, 0.18)
Mediana (1st Qu, 3rd Qu, IQR)
12.7 (11.5, 14.12, 2.62)
*****
Descriptiva hb_fin
*****
Número de elementos: 96
NAs: 0
Media (sd, se)
12.84 (1.615, 0.16)
Mediana (1st Qu, 3rd Qu, IQR)
12.75 (11.6, 13.82, 2.22)
*****
TEST DE LILLIEFORS para hb0
p-valor: 0.3817
*****
TEST DE LILLIEFORS para hb_fin
p-valor: 0.7309
*****
***** TEST PARAMÉTRICOS EN POBLACIONES NORMALES
*****
*****
***** Comparación de medias para muestras relacionadas
*****

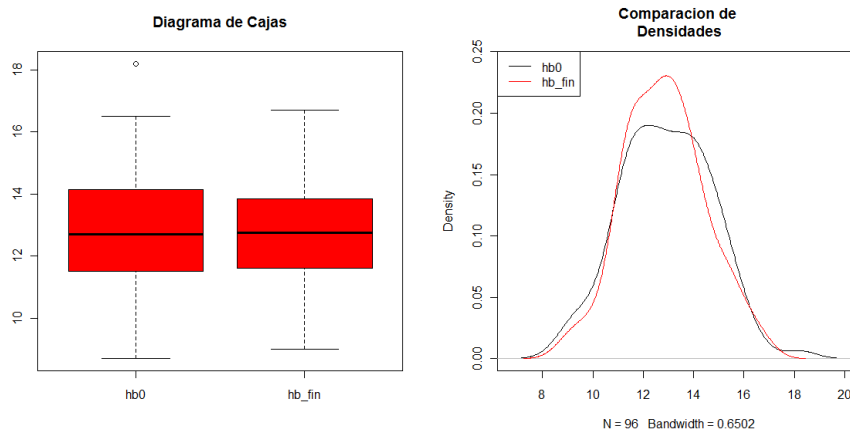
    Paired t-test

data: covar1.var and covar2.var
t = 0.4605, df = 95, p-value = 0.6462
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.2311016  0.3706850
sample estimates:
mean of the differences
      0.06979167

TIENEN LA MISMA MEDIA

```

Figura.3 Diagrama de cajas y las diferentes densidades de hemoglobina basal y final



Capítulo 4

Evaluación del impacto de una consulta especializada en el servicio de medicina interna.

1. INTRODUCCIÓN

La insuficiencia cardíaca (Heart Failure, HF) es la incapacidad del corazón de bombear sangre en los volúmenes más adecuados para satisfacer las demandas del organismo.

La HF constituye un problema de salud pública importante, y su incidencia ha aumentado en la última década coincidiendo con el envejecimiento de la población. Su prevalencia afecta al 10% de los individuos que superan los 75 años, y aumenta con la edad. Constituye la primera causa de hospitalización y reingreso en mayores de 65 años, suponiendo la mayor proporción del gasto sanitario. Además, la mortalidad es alta (siendo la tercera de las causas cardiovasculares por detrás de la cardiopatía isquémica y la enfermedad vascular cerebral) y merma de forma considerable la calidad de vida de los pacientes que la sufren.

En el año 2007 el servicio de Medicina Interna del Complejo Hospitalario Universitario de Santiago de Compostela (CHUS) pone en marcha una consulta especializada para el seguimiento y control de los pacientes con insuficiencia cardíaca. La hipótesis de la que parten los investigadores es que un mejor control y seguimiento de los pacientes con HF permite reducir su mortalidad. Para lograr el objetivo del estudio y en ausencia de grupo control definido a priori, en este trabajo se consideró como grupo control aquellos individuos que asisten al menos a la segunda consulta.

Los objetivos de este estudio son los siguientes:

- Caracterizar a los individuos controlados en la consulta especializada frente a los individuos que no se controlaron.
- Caracterizar a los individuos controlados, según el motivo de finalización del seguimiento en la consulta especializada (muerte, derivación a otro servicio, o abandono del estudio por otras causas) y definir los factores pronóstico de estos motivos de finalización del seguimiento.

- Describir las causas de muerte (cardiovasculares y no cardiovasculares) y definir los factores pronóstico de mortalidad.
- Describir la repercusión del seguimiento de los pacientes controlados hasta la fecha fin de seguimiento (30/06/2011).
- Valorar la supervivencia de una cohorte de pacientes con insuficiencia cardíaca en una consulta de Medicina Interna durante un período de 4 años y medio.

2. MÉTODOS

2.1. Diseño del estudio

Estudio de cohortes retrospectivo, no intervencionista, de pacientes diagnosticados de HF seguidos en una consulta específica de Medicina Interna del Complejo Hospitalario Universitario de Santiago de Compostela (CHUS).

El seguimiento en la consulta especializada, se hace de modo protocolizado, siguiendo las directrices de la Guía de Insuficiencia Cardíaca de la Sociedad Española de Cardiología (SEC).

2.2. Sujetos

En este estudio 277 pacientes con insuficiencia cardíaca asistieron, al menos una vez, a una consulta específica del servicio de Medicina Interna del CHUS, en el periodo comprendido entre el 01/01/2007 hasta el 31/12/2010. De los 277 pacientes, 22 asistieron solamente a la primera consulta y abandonaron el estudio por causas ajenas a la HF (grupo no controlados) y 255 asistieron al menos a la segunda consulta (grupo controlados). En este último grupo y en relación con el mantenimiento de su adherencia a la consulta específica: se mantuvieron en consulta 96 pacientes hasta la fecha fin de estudio (30/06/2011), 71 pacientes abandonaron con anterioridad la consulta por causas ajenas a la HF (pudiendo estar vivos o muertos a la fecha fin de estudio), 34 pacientes fallecieron estando en consulta (fecha de muerte anterior a la fecha fin de estudio) y a 54 pacientes se les dio el alta y se derivaron a otros servicios (pudiendo estar vivos o muertos a la fecha fin de estudio).

2.3. Variables

Variables demográficas:

- *edad_entrada*: número de años que tiene el paciente cuando asiste a su primera consulta.
- *sexo*: si es un hombre (se codifica con 1) y si es una mujer (se codifica con 2).

VARIABLES DE COMORBILIDAD

Son enfermedades crónicas presentes en cada sujeto:

- *hta*: hipertensión arterial (1=si, 0=no).
- *dm*: diabetes mellitus (1=si, 0=no).
- *epoc*: enfermedad pulmonar obstructiva crónica (1=si, 0=no).
- *fa*: fibrilación auricular (1=si, 0=no).

VARIABLES DE TIEMPO:

- *consulta1*: es la fecha correspondiente a la primera consulta realizada por cada paciente. Esta primera consulta se puede realizar desde el 01/01/2007 hasta el 31/12/2010.
- *fecha_final*: si el paciente está vivo, esta fecha es 30/06/2011, y si el paciente falleció la fecha es anterior a la fecha fin de estudio (30/06/2011).
- *tiempo_dias*: es el tiempo de supervivencia (en días) desde que la persona entra en el estudio (*consulta1*) hasta la *fecha_final*.

VARIABLES DE CONTROL:

- *status*: si el sujeto está muerto (exitus=1) o si está vivo (censura=0) al finalizar el estudio.
- *n_cons_seg*: es el número total de consultas que realiza el sujeto.
- *motivo*: los motivos del fin de seguimiento, de los 255 pacientes que se controlaron, son: seguido=1, pérdida=2, muerte=3 y alta=4.
- *grupo*: si el paciente solo asiste a la primera consulta (no controlado=0) y si asiste al menos a la segunda consulta (controlado=1).

VARIABLES BASEALES Y FINALES

Las variables que indican la situación clínica del paciente a su inclusión en el estudio son las basales, y las que indican la situación clínica al final del control en consulta de HF son las finales.

- *glucemia*: *glucemia0* (basal) y *glucemia_fin* (final). Es la concentración de glucosa (mg/dl) en sangre.

- *sodio*: *sodio0* (basal) y *sodio_fin* (final). Es la concentración de sodio (mMol/l) en sangre.
- *hiponatremia*: *hiponatremia0* (basal) e *hiponatremia_fin* (final). Tener hiponatremia es tener un nivel de sodio en sangre por debajo de los límites de normalidad. (hiponatremia=1, no hiponatremia=0).
- *GFR*: *GFR0* (basal) y *GFR_fin* (final). Es la tasa de filtrado glomerular calculada a partir de la edad, creatinina y sexo usando la fórmula MDRD; es el indicador más utilizado para valorar la función renal. (Glomerular Filtration Rate, GFR).
- *ERC*: *ERC0* (basal) y *ERC_fin* (final). Presencia de Enfermedad Renal Crónica, definida como un GFR<60 según las Guías KDQI. (ERC=1, no ERC=0).
- *hemoglobina*: *hb0* (basal) y *hb_fin* (final). Es la concentración de hemoglobina (g/dl) en sangre.
- *anemia*: *anemia0* (basal) y *anemia_fin* (final). La anemia se define como una concentración de hemoglobina en la sangre inferior a 11 g/dl. (anemia=1, no anemia=0).
- *NYHA*: *NYHA0* (basal) y *NYHA_fin* (final). La NYHA (New York Heart Association) es la clasificación más utilizada para establecer la capacidad funcional de pacientes con insuficiencia cardíaca. Establece cuatro clases (I, II, III, IV), basándose en las limitaciones en la actividad física del paciente ocasionadas por los síntomas cardíacos, y son las siguientes:
 - Clase I. No limitación de la actividad física. La actividad ordinaria no ocasiona excesiva fatiga, palpitaciones, disnea o dolor anginoso.
 - Clase II. Ligera limitación de la actividad física. Confortables en reposo. La actividad ordinaria ocasiona fatiga, palpitaciones, disnea o dolor anginoso.
 - Clase III. Marcada limitación de la actividad física. Confortables en reposo. Actividad física menor que la ordinaria ocasiona fatiga, palpitaciones, disnea o dolor anginoso.
 - Clase IV. Incapacidad para llevar a cabo cualquier actividad física sin disconfort. Los síntomas de insuficiencia cardíaca o de síndrome anginoso pueden estar presentes incluso en reposo. Si se realiza cualquier actividad física, el disconfort aumenta.

En nuestro estudio codificamos NYHA=1, agrupando las clases I y II, y NYHA=3, agrupando las clases III y IV.

3. ANÁLISIS ESTADÍSTICO

Cada una de las variables: demográficas, de comorbilidad y basales, se definen como las variables del estudio. A continuación, nombramos las técnicas estadísticas utilizadas para conseguir los siguientes objetivos:

- a) Para caracterizar a los individuos controlados en la consulta especializada frente a los que no se controlaron, se comprueba la asociación entre cada una de las variables del estudio y el *grupo* (control, no control), usando:
 - El t-test y el test de Wilcoxon para variables continuas.
 - El test chi-cuadrado y el test exacto de Fisher para variables cualitativas.
- b) Para describir las causas de muerte (cardiovasculares y no cardiovasculares), se comprueba la asociación entre cada una de las variables del estudio y el *status* (muerto o vivo) usando:
 - El t-test y el test de Wilcoxon para variables continuas.
 - El test chi-cuadrado para variables cualitativas.

Con la finalidad de evaluar los factores pronóstico de mortalidad, se utiliza un modelo de regresión con respuesta binaria (*status*).

- c) Para caracterizar a los individuos controlados, según el motivo de finalización del seguimiento en la consulta especializada, se comprueba la asociación entre cada una de las variables del estudio y el *motivo* (seguido, pérdida, muerte y alta), usando:
 - ANOVA de un factor y el test de Kruskal-Wallis para variables continuas.
 - El test chi-cuadrado para variables cualitativas.

Con el objetivo de establecer los factores pronóstico de estos motivos de finalización del seguimiento, se utiliza un modelo de regresión con respuesta multinomial (*motivo*).

- d) Para describir la repercusión del seguimiento en la consulta específica, de los 96 pacientes que fueron seguidos hasta la fecha fin de estudio (30/12/2010), se comparan las variables basales y finales de esos pacientes, para comprobar si hubo cambios a lo largo de ese seguimiento, usando:
 - El t-test y el test de Wilcoxon para muestras relacionadas, si las variables son continuas.

- El test de Mc-Nemar para muestras relacionadas, si las variables son cualitativas.
- e) Valorar la supervivencia de una cohorte de pacientes con insuficiencia cardíaca en una consulta de Medicina Interna durante un período de 4 años y medio, usando:
- El estimador de Kaplan-Meier.
 - Un modelo de regresión de Cox.

3.1. TEST PARAMÉTRICOS EN DOS POBLACIONES NORMALES

Prueba F de comparación de varianzas

Sea (X_1, X_2, \dots, X_n) una muestra de tamaño n procedente de una distribución normal $X \approx N(\mu_X, \sigma_X)$ y sea (Y_1, Y_2, \dots, Y_m) una muestra de tamaño m de otra población normal $Y \approx N(\mu_Y, \sigma_Y)$. Para contrastar $H_0 : \sigma_X^2 = \sigma_Y^2$ ($\sigma_X^2 / \sigma_Y^2 = 1$) se utiliza el estadístico de contraste:

$$E = \frac{S_X^2}{S_Y^2} = \left(\frac{(n-1)S_X^2 / \sigma_X^2 (n-1)}{(m-1)S_Y^2 / \sigma_Y^2 (m-1)} \right) \approx F_{n-1, m-1}, \text{ si } H_0 \text{ es cierta.}$$

Se rechaza H_0 si el p-valor calculado para el valor de E en la muestra es menor que α .

Prueba t de comparación de medias para muestras independientes

Sea (X_1, X_2, \dots, X_n) una muestra de tamaño n procedente de una distribución normal, $X \approx N(\mu_X, \sigma_X)$ y sea (Y_1, Y_2, \dots, Y_m) una muestra de tamaño m de otra población normal $Y \approx N(\mu_Y, \sigma_Y)$, independiente de la anterior. Para contrastar la igualdad de medias,

$H_0 : \mu_X - \mu_Y = 0$ vs $H_1 : \mu_X - \mu_Y \neq 0$, el estadístico cambia dependiendo de que se asuman varianzas poblaciones iguales o distintas.

$$E = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\hat{\sigma}(\bar{X} - \bar{Y})} \approx t_v, \text{ si } H_0 \text{ es cierta}$$

En el caso de varianzas iguales $v = n + m - 2$.

En el caso de varianzas distintas v , además de depender de n y m , depende de las varianzas muestrales: $S_X^2 = \sum_{i=1}^n \frac{(X_i - \bar{X})^2}{n-1}$ y $S_Y^2 = \sum_{j=1}^m \frac{(Y_j - \bar{Y})^2}{m-1}$. Para el valor de E en la muestra, se calcula su p-valor y se rechaza si es menor que α .

Prueba t de comparación de medias para muestras dependientes.

En este caso se considera una muestra (X_1, X_2, \dots, X_n) de tamaño n procedente de una distribución normal $X \approx N(\mu_X, \sigma_X)$ y una muestra (Y_1, Y_2, \dots, Y_n) de tamaño también n , de otra población normal $Y \approx N(\mu_Y, \sigma_Y)$. Se pueden reducir los datos a una sola muestra (D_1, D_2, \dots, D_n) , donde cada $D_i = X_i - Y_i$, tendrá también distribución normal con media $\mu_D = \mu_X - \mu_Y$. Este caso es habitual cuando se toman medidas repetidas en los individuos de una población.

Para contrastar $H_0 : \mu_X - \mu_Y = 0$ se utiliza el estadístico de contraste

$$E = \frac{\bar{D} - \mu_D}{S_D / \sqrt{n}} \approx t_{n-1}, \text{ si } H_0 \text{ es cierta.}$$

Se rechaza H_0 si el p-valor calculado para el valor de E en la muestra es menor que α .

3.2. TEST PARAMÉTRICO EN MÁS DE DOS POBLACIONES NORMALES

Anova de un factor

Además de que las $k > 2$ variables sean normales tiene que cumplirse que las k varianzas sean homogéneas, esto se evalúa realizando el Test de Levene (Levene, H. (1960)).

En un análisis de varianza (ANOVA) de un factor el objetivo que se busca es evaluar si los valores medios de una variable continua varían en función de las categorías, en nuestro caso los diferentes motivos del fin de seguimiento, de una variable cualitativa. Esta prueba es la extensión de la prueba t para más de dos categorías. Queremos contrastar:

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_k = \mu \quad \text{vs} \quad H_1 : \mu_i \neq \mu_j \text{ para algún } i \text{ e } j.$$

La hipótesis nula nos indica que no hay diferencias en las medias en los distintos motivos y la alternativa nos indica que al menos las medias en dos de los motivos no son iguales. Para contrastar la hipótesis anterior contaremos con

una muestra (o $k = 4$ muestras si consideramos por separado cada uno de los motivos), que representaremos de la siguiente manera:

Motivo	Muestra	Tamaño de muestra	Media
1	$Y_{11}, Y_{12}, \dots, Y_{1n_1}$	n_1	$\bar{Y}_1 = \frac{\sum_{j=1}^{n_1} Y_{1j}}{n_1}$
2	$Y_{21}, Y_{22}, \dots, Y_{2n_2}$	n_2	$\bar{Y}_2 = \frac{\sum_{j=1}^{n_2} Y_{2j}}{n_2}$
3	$Y_{31}, Y_{32}, \dots, Y_{3n_3}$	n_3	$\bar{Y}_3 = \frac{\sum_{j=1}^{n_3} Y_{3j}}{n_3}$
4	$Y_{41}, Y_{42}, \dots, Y_{4n_4}$	n_4	$\bar{Y}_4 = \frac{\sum_{j=1}^{n_4} Y_{4j}}{n_4}$
Global		$n = n_1 + n_2 + n_3 + n_4$	$\bar{Y}_{..} = \frac{\sum_{i=1}^4 \sum_{j=1}^{n_i} Y_{ij}}{n}$

Cada una de las observaciones de nuestra muestra Y_{ij} puede ser representada de la siguiente manera:

$$Y_{ij} = \mu_i + (Y_{ij} - \mu_i) = \mu_i + e_{ij},$$

donde e_{ij} representa el error de la observación j del motivo i con respecto a la media en ese motivo. Además,

$$Y_{ij} = \mu_i + e_{ij} = \mu + (\mu_i - \mu) + e_{ij} = \mu + \alpha_i + e_{ij}$$

Por lo tanto cada observación la podemos descomponer en:

1. μ , que es la media global
2. α_i , que es la desviación entre la media global y la media en el motivo i (efecto del motivo i).
3. e_{ij} , que es el error experimental.

El estadístico de contraste en el ANOVA de un factor, se basa en comparar la variabilidad entre motivos (diferencias en las medias) y la variabilidad de la respuesta dentro de cada motivo (σ^2).

$$\frac{\text{Variabilidad entre motivos}}{\text{Variabilidad dentro de los motivos}}$$

Si el cociente anterior es “grande”, entonces cabe esperar que sí existan diferencias entre los motivos.

Teniendo en cuenta que,

$$Y_{ij} = \mu + \alpha_i + e_{ij} = \mu + (\mu_i - \mu) + (Y_{ij} - \mu_i),$$

substituimos los correspondientes valores poblacionales μ_i y μ por los correspondientes valores muestrales,

$$Y_{ij} = \bar{Y}_{..} + (\bar{Y}_i - \bar{Y}_{..}) + (Y_{ij} - \bar{Y}_i),$$

llegando a la siguiente igualdad,

$$(Y_{ij} - \bar{Y}_{..}) = (\bar{Y}_i - \bar{Y}_{..}) + (Y_{ij} - \bar{Y}_i).$$

A partir de esta última expresión, se obtiene la siguiente **suma de cuadrados**,

$$\sum_{i=1}^4 \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{..})^2 = \sum_{i=1}^4 n_i (\bar{Y}_i - \bar{Y}_{..})^2 + \sum_{i=1}^4 \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2.$$

En la igualdad anterior, podemos identificar 3 componentes:

1. Variabilidad total o suma total de cuadrados

$$SC_{Total} = \sum_{i=1}^4 \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{..})^2.$$

Medida de la variabilidad de los datos con respecto a la media global.

2. Variabilidad explicada o suma de cuadrados de los motivos

$$SC_{Motivo} = \sum_{i=1}^4 n_i (\bar{Y}_i - \bar{Y}_{..})^2.$$

Medida de la variabilidad de las medias en cada motivo con respecto a la media global. (Variabilidad entre motivos).

3. Variabilidad no explicada o suma de cuadrados del error

$$SC_{Error} = \sum_{i=1}^4 \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})^2.$$

Medida de la variabilidad en la respuesta entre observaciones del mismo motivo. (Variabilidad dentro de cada tratamiento).

Por lo que el estadístico de contraste es de la siguiente manera:

$$F_0 = \frac{\frac{SC_{Motivo}}{3}}{\frac{SC_{Error}}{n-4}} \approx F_{3,n-4}, \text{ si } H_0 \text{ es cierta.}$$

Se rechaza H_0 si el p-valor calculado para el valor de F_0 en la muestra es menor que α . En este caso el objetivo será evaluar que motivos son distintos y que motivos son iguales. En términos de un contraste, lo que buscamos es:

$$H_0 : \mu_i = \mu_j \text{ vs } H_1 : \mu_i \neq \mu_j, \text{ para cada par de motivos } i \text{ e } j.$$

Existen diferentes métodos para realizar el anterior contraste, entre ellos:

- Método de la mínima diferencia significativa.
- Método de Bonferroni.
- Método de la diferencia significativa de Tukey.

3.3. TEST NO PARAMÉTRICO EN DOS MUESTRAS

Test de Wilcoxon para muestras independientes

Sea (X_1, X_2, \dots, X_n) una muestra de tamaño n procedente de una población con distribución de tipo continuo F_X y sea (Y_1, Y_2, \dots, Y_m) una muestra de tamaño m de otra población con distribución de tipo continuo F_Y .

Se trata de contrastar $H_0 : F_X = F_Y$ vs $H_1 : F_X \neq F_Y$. Este test se basa en los rangos de las observaciones y utiliza la idea de que, si H_0 es cierta, se espera que los rangos correspondientes a los valores de una y otra muestra estén entremezclados o dispersos, mientras que en otro caso, debe esperarse que los rangos de las observaciones de cada muestra estén muy agrupados en los extremos.

Llamando R a la suma de los rangos asociados a las observaciones de una cualquiera de las dos muestras (por ejemplo, la de menor tamaño), la distribución del estadístico R está tabulada. El estadístico R está perfectamente relacionado con el estadístico U de Mann-Whitney, de manera que ambos proporcionan la misma prueba.

3.3.1. Test de Wilcoxon para muestras dependientes

La hipótesis a contrastar es la misma que la del anterior, $H_0 : F_X = F$, pero en este caso se considera una muestra de tamaño n de una población bivalente, es decir n pares de observaciones muestrales $(X_1, Y_1), \dots, (X_n, Y_n)$ correspondientes a dos variables aleatorias X e Y con distribución de tipo continuo, pero no necesariamente independientes.

Como ya se comentó anteriormente, esta situación aparece frecuentemente al tomar medidas repetidas sobre un mismo individuo.

Si reducimos los datos a una sola muestra D_1, D_2, \dots, D_n , donde cada $D_i = X_i - Y_i$, pasamos a trabajar con una muestra de tamaño n de la variable aleatoria $D = X - Y$.

Bajo la hipótesis nula, esta variable diferencia estará distribuida de forma simétrica respecto al 0, por lo que las diferencias positivas y negativas de igual magnitud absoluta serán equiprobables. Los estadísticos de Wilcoxon se basan en esta idea y calculan la suma de los rangos de las diferencias positivas, $T+$, y de las diferencias negativas, $T-$, habiéndose asignado los rangos a los valores absolutos de las diferencias observadas.

La distribución de los estadísticos de Wilcoxon (por ejemplo del $T+$) está tabulada.

3.4. TEST NO PARAMÉTRICO PARA MÁS DE DOS MUESTRAS

Test de Kruskal-Wallis para k muestras independientes

Se trata de una generalización del test de Wilcoxon al caso de k muestras. Sean $(X_{11}, X_{12}, \dots, X_{1n_1}), (X_{21}, X_{22}, \dots, X_{2n_2}), \dots, (X_{k1}, X_{k2}, \dots, X_{kn_k})$, k muestras de tamaños n_1, n_2, \dots, n_k , respectivamente, de sendas distribuciones continuas.

Se quiere contrastar la hipótesis nula de que las k distribuciones son idénticas frente a la alternativa de que difieren en su tendencia central.

A cada una de las $N = n_1 + n_2 + \dots + n_k$ observaciones conjuntamente ordenadas, se les asigna un rango y se determina la suma de los rangos R_j de las observaciones de cada una de las muestras, $j = 1, \dots, k$.

En esencia, el contraste determina si la disparidad entre las R_j con respecto a los tamaños muestrales n_j de las muestras es suficientemente significativa para llevar al rechazo de la hipótesis nula.

Si H_0 es cierta y para tamaños muestrales suficientemente grandes, el estadístico

$$H = \frac{12}{N(N+1)} \left(\sum_{j=1}^k \left(\frac{R_j}{n_j} \right) \right) - 3(N+1) \approx \chi_{k-1}^2,$$

Por lo que para un nivel de significación α , se rechazará H_0 si el estadístico $H >$ el percentil $1 - \alpha$ de una χ_{k-1}^2 .

3.5. INDEPENDENCIA EN TABLAS DE CONTINGENCIA

Nuestro objetivo será contrastar la hipótesis de independencia poblacional entre dos factores basándonos en la información proporcionada por las frecuencias observadas contenidas en la tabla de contingencia.

Tabla.1. (Notación para una tabla de contingencia $I \times J$)

		Factor B					
		B_1	...	B_1	...	B_j	
Factor A	A_1	n_{11}	...	n_{1j}	...	n_{1J}	$n_{1.}$

	A_i	n_{i1}	...	n_{ij}	...	n_{iJ}	$n_{i.}$

	A_I	n_{I1}	...	n_{Ij}	...	n_{IJ}	$n_{I.}$
		$n_{.1}$...	$n_{.j}$...	$n_{.J}$	n

A partir de la Tabla.1 se obtienen las siguientes distribuciones de frecuencias marginales y condicionales:

Distribuciones marginales:

Factor A:
$$n_{i.} = \sum_{j=1}^J n_{ij} \quad (i = 1, \dots, I),$$

Factor B:
$$n_{.j} = \sum_{i=1}^I n_{ij} \quad (j=1, \dots, J),$$

verificando
$$n = \sum_{i=1}^I \sum_{j=1}^J n_{ij} = \sum_{i=1}^I n_{i.} = \sum_{j=1}^J n_{.j}.$$

Distribuciones condicionadas:

Factor A condicionado a $B = B_j$ ($j = 1, \dots, J$):

$$n_{i|j} = n_{ij} / n_{.j} \quad i = 1, \dots, I$$

Factor B condicionado a $A = A_i$ ($i = 1, \dots, I$):

$$n_{j|i} = n_{ij} / n_{i.} \quad j = 1, \dots, J$$

Tabla.2. (Notación para una tabla $I \times J$ de probabilidades poblacionales)

		Factor B					
		B_1	...	B_1	...	B_J	
Factor A	A_1	p_{11}	...	p_{1j}	...	p_{1J}	$p_{1.}$

	A_i	p_{i1}	...	p_{ij}	...	p_{iJ}	$p_{i.}$

	A_I	p_{I1}	...	p_{Ij}	...	p_{IJ}	$p_{I.}$
		$p_{.1}$...	$p_{.j}$...	$p_{.J}$	1

De igual forma que para las frecuencias observadas, se definen las distribuciones de probabilidad marginales y condicionadas asociadas de la siguiente manera:

Distribuciones marginales:

Factor A:
$$p_{i.} = \sum_{j=1}^J p_{ij} \quad (i = 1, \dots, I),$$

Factor B:
$$p_{.j} = \sum_{i=1}^I p_{ij} \quad (j = 1, \dots, J),$$

verificando
$$\sum_{i=1}^I p_{i.} = \sum_{j=1}^J p_{.j} = 1$$

Distribuciones condicionadas:

Factor A condicionado a $B = B_j$ ($j = 1, \dots, J$):

$$p_{i|j} = p_{ij} / p_{.j} \quad i = 1, \dots, I$$

Factor B condicionado a $A = A_i$ ($i = 1, \dots, I$):

$$p_{j|i} = p_{ij} / p_{i.} \quad j = 1, \dots, J$$

Contraste de independencia Chi-cuadrado.

Consideremos una tabla de contingencia $I \times J$ generada por muestreo multinomial. Esto significa que la distribución conjunta del vector aleatorio que genera a las frecuencias observadas es una multinomial de parámetros las probabilidades poblacionales.

Supongamos que queremos contrastar la hipótesis nula de independencia

$$H_0 : p_{ij} = p_{i.} p_{.j} \quad \forall i = 1, \dots, I ; j = 1, \dots, J$$

Sean $\hat{m}_{ij} = n_i n_{.j} / n$ los estimadores de máxima verosimilitud de las frecuencias esperadas bajo la hipótesis de independencia. Para llevar a cabo este contraste, se define el estadístico de Pearson de la siguiente manera:

$$X^2 = \sum_i \sum_j \frac{(n_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}} ,$$

que bajo la hipótesis nula de independencia, tiene distribución de probabilidad asintótica (para tamaños muestrales grandes) $\chi^2_{(I-1)(J-1)}$.

Se rechaza la hipótesis de independencia al nivel α cuando $X^2 \geq \chi^2_{(I-1)(J-1); \alpha}$

Corrección por continuidad del estadístico X^2

Los contrastes de independencia basados en la distribución chi-cuadrado aproximan una distribución discreta (la del estadístico del contraste que es función de la multinomial) por una continua (la distribución chi-cuadrado).

Esto da lugar a un error considerable cuando el tamaño muestral no es demasiado grande. Los métodos que disminuyen este error se llaman correcciones por continuidad.

Para mejorar esta aproximación en el caso de tamaños muestrales pequeños, Yates (Yates, 1934) propuso un método que se llama corrección por continuidad de Yates para el estadístico X^2 en una tabla 2×2 .

Esta corrección consiste en restar $1/2$ a las desviaciones positivas y sumar $1/2$ a las desviaciones negativas. Por lo tanto, el estadístico X^2 corregido es de la forma:

$$X_C^2 = \sum_i \sum_j \frac{(|n_{ij} - \hat{m}_{ij}| - (1/2))^2}{\hat{m}_{ij}}$$

Existe mucha polémica sobre la aplicación de la corrección de Yates porque hay casos prácticos en los que con el estadístico corregido se rechaza la independencia con bastante menor significación que con el test no corregido. Es decir, el p-valor corregido es bastante mayor que el no corregido y el test corregido es más conservativo. En caso de tamaño muestral grande la corrección tiene efecto pequeño sobre el valor del estadístico X^2 .

Test exacto de Fisher para tablas 2x2

Consideremos una tabla 2×2 con los totales marginales $(n_{1.}, n_{2.}, n_{.1}, n_{.2})$ de ambas variables fijos.

Supongamos que queremos contrastar la hipótesis de independencia poblacional que para una tabla 2×2 se puede expresar como:

$$H_0 : p_{11} = p_{12}$$

La distribución exacta bajo H_0 de cualquier distribución de frecuencias observadas cuyos totales marginales coincidan con los fijados de antemano, es la hipergeométrica de parámetros $(n, n_{1.}, n_{.1})$ dada por

$$\frac{\binom{n_{1.}}{n_{11}} \binom{n - n_{1.}}{n_{.1} - n_{11}}}{\binom{n}{n_{.1}}} = \frac{n_{1.}! n_{2.}! n_{.1}! n_{.2}!}{n! n_{11}! n_{12}! n_{21}! n_{22}!}$$

que bajo la hipótesis de independencia no depende de ningún parámetro desconocido, y expresa la distribución de las cuatro casillas de la tabla en términos del elemento n_{11} debido a que, dados los totales marginales, su valor determina las otras tres casillas de la tabla.

El rango para los posibles valores de n_{11} $\max\{0, n_{.1} - (n - n_{1.})\} \leq n_{11} \leq \min\{n_{1.}, n_{.1}\}$.

Dado que aparecen en la fórmula de la distribución hipergeométrica un total de $IJ + I + J + 1$ factoriales, el cálculo de las probabilidades de todas las posibles tablas

es complicado y laborioso. Este proceso se simplifica aplicando la fórmula de Feldman y Kinger que calcula una de estas probabilidades, por ejemplo la de la tabla observada, y las demás se obtienen a partir de ella. Si denotamos por $p_{n_{11}}$ a la probabilidad que asigna la distribución hipergeométrica a una tabla con frecuencia n_{11} en la casilla (1,1), la expresión para la probabilidad de las demás tablas es:

$$p_{n_{11}+1} = \frac{n_{12}n_{21}}{(n_{11}+1)(n_{22}+1)} p_{n_{11}},$$

$$p_{n_{11}-1} = \frac{n_{11}n_{22}}{(n_{12}+1)(n_{21}+1)} p_{n_{11}}.$$

El paso siguiente es fijar una hipótesis alternativa, fijemos $H_1 : p_{11} \neq p_{12}$. Como bajo H_0 se tiene que verificar $n_{11} = \hat{m}_{11} = n_{1.}n_{.1}/n$, se rechazará H_0 cuando la distancia $|d_{11}| = |n_{11} - \hat{m}_{11}|$ sea suficientemente grande. Por lo tanto el p-valor del test es:

$$p = P[n_{11} : |d_{11}| \geq |d_{11Obs}|] \text{ definiendo } d_{11Obs} = n_{11Obs} - \hat{m}_{11}.$$

Si $n_{11Obs} \geq \hat{m}_{11}$ entonces $p = \sum_{\{n_{11}: n_{11} \in (-n_{11Obs} + 2\hat{m}_{11}, n_{11Obs})\}} p_{n_{11}}.$

Si $n_{11Obs} < \hat{m}_{11}$ entonces $p = \sum_{\{n_{11}: n_{11} \in (n_{11Obs}, -n_{11Obs} + 2\hat{m}_{11})\}} p_{n_{11}}.$

El test resulta significativo para rechazar la independencia si $p \leq \alpha$ siendo α el nivel de significación previamente fijado para el contraste.

Para más información sobre este test puede consultarse (Aguilera, 2001).

Los test vistos anteriormente sirven para contrastar si dos variables cualitativas son independientes, pero en ocasiones tenemos variables cualitativas tomadas de un mismo individuo, por ejemplo en nuestro caso *anemia0* (basal) y *anemia_fin* (final), por lo que el estudio de la independencia no tiene interés. Las tablas que resultan de la representación de estos datos (relacionados o dependientes) son cuadradas y además con las mismas categorías para filas y columnas.

Las hipótesis de interés en este tipo de tablas serán (debido a su interpretación en relación a si existen cambios significativos de la variable de interés en el tiempo o bajo los tratamientos considerados): la simetría y la homogeneidad marginal.

- Simetría: $p_{ij} = p_{ji} \quad (i \neq j)$
- Homogeneidad marginal: $p_{.i} = p_{.i} \quad (i = 1, \dots, I)$

En una tabla 2×2 ambos conceptos son equivalentes. Sin embargo, en general, la simetría implica homogeneidad marginal pero el recíproco no es cierto.

Test de Mc-Nemar para una tabla 2x2.

La estructura de una tabla 2×2 procedente de datos dependientes binarios (clasificados según una variable dicotómica) es la siguiente:

	Caso2	
Caso1	A1	A2
A1	n_{11}	n_{12}
A2	n_{21}	n_{22}

Contiene en cada una de sus casillas las frecuencias observadas n_{ij} .

Vamos a obtener en primer lugar un test aproximado (basado en la distribución chi-cuadrado) para contrastar la hipótesis nula:

$$H_0 : p_{1.} = p_{.1} \Leftrightarrow p_{12} = p_{21} \quad \text{vs} \quad H_1 : p_{1.} \neq p_{.1} \Leftrightarrow p_{12} \neq p_{21}$$

H_0 significa que la probabilidad de clasificarse en la categoría A1 es la misma en los dos casos considerados. Es decir, se contrasta que no hay cambios significativos de la clasificación en los dos casos frente a la existencia de cambios. Por ello el test se suele llamar, también, test de significación de cambios.

El test es debido a McNemar (McNemar, 1947) y se obtiene como sigue. Bajo H_0 los EMV (suponiendo muestreo multinomial para la tabla 2×2) de las probabilidades poblacionales son:

$$\hat{p}_{11} = n_{11} / n \quad \hat{p}_{22} = n_{22} / n \quad \hat{p}_{12} = \hat{p}_{21} = (n_{12} + n_{21}) / 2n$$

Por lo tanto, los EMV de las frecuencias esperadas bajo H_0 son:

$$\hat{m}_{11} = n_{11} \quad \hat{m}_{22} = n_{22} \quad \hat{m}_{12} = \hat{m}_{21} = (n_{12} + n_{21}) / 2$$

Entonces, substituyendo en el estadístico X^2 de bondad de ajuste se tiene la expresión:

$$X^2 = \frac{(n_{12} - n_{21})^2}{(n_{12} + n_{21})}$$

que tiene, bajo H_0 distribución asintótica χ^2 con 1 grado de libertad, y es conocido como estadístico de McNemar.

El número de parámetros libres de la tabla es 3. El número de parámetros estimados es dos. Por lo tanto la diferencia es el número de grados de libertad igual a 1.

Por lo tanto, el test resultará significativo para rechazar la hipótesis de independencia al nivel α si se verifica: $X^2 \geq \chi_{1,\alpha}^2$

Se puede comprobar que la corrección de Yates del estadístico de McNemar es de la forma:

$$X_C^2 = \frac{(|n_{12} - n_{21}| - 1)^2}{(n_{12} + n_{21})}$$

La corrección de Yates es aconsejable cuando las frecuencias esperadas en el caso de simetría (u homogeneidad marginal) son menores que 5, es decir, $\frac{(n_{12} + n_{21})}{2} < 5$.

En nuestro estudio no tenemos variables cualitativas dependientes con más de dos categorías, si las tuviésemos para contrastar la simetría tendríamos que utilizar la extensión de Bowker (Bowker, 1948) para al test de McNemar. Si la simetría se rechaza, como no estamos en tablas 2x2 la simetría y la homogeneidad marginal no son equivalentes por lo que puede ser de interés utilizar la extensión de Stuart (Stuart, 1955) y Maxwell (Maxwell, 1970) al test de McNemar para contrastar la homogeneidad marginal.

3.6. REGRESIÓN LOGÍSTICA

Consideremos una variable respuesta Y , que sólo puede tomar dos valores, que codificaremos como 0 y 1. Diremos que Y es una variable binaria (o dicotómica), y queremos estudiar su relación con otras variables explicativas X mediante un análisis de regresión. Por ser la variable respuesta binaria, su distribución será Bernoulli y su media una probabilidad de éxito.

Nuestro propósito es construir un modelo para

$$\pi(x) = P(Y = 1 / X = x)$$

esto es, para la probabilidad de éxito condicionada a cada valor de la variable explicativa. Podemos aplicar a $\pi(x)$ una función que transforme el intervalo $[0,1]$ en toda la recta real. Así, se trataría de expresar

$$g(\pi(x, \beta)) = x' \beta$$

donde a la función g se le conoce como función enlace o función link, x es el vector que contiene los valores de las variables explicativas y β el vector de los coeficientes.

En esta situación donde la variable respuesta es dicotómica, es común considerar como función link la función logística o función logit:

$$g(p) = \frac{p}{1-p} \quad \forall p \in [0,1]$$

Como el argumento p será sustituido por la probabilidad de éxito, la función logística consiste en efectuar un logaritmo al cociente entre la probabilidad de éxito, p , y la probabilidad de fracaso, $(1-p)$. A este cociente se le conoce como odds:

$$Odds(Y) = \frac{P(Y=1)}{P(Y=0)}$$

Ahora que la odds se mueve en el intervalo $[0, +\infty]$, sólo falta aplicar un logaritmo para transformarlo en una cantidad situada en toda la recta real. En definitiva el modelo consistirá en expresar el logaritmo de la odds de la variable respuesta como función lineal de la variable explicativa:

$$\log \frac{\pi(x, \beta)}{1 - \pi(x, \beta)} = x' \beta$$

Veamos ahora que forma adopta este modelo si lo pensamos como representación de la probabilidad de éxito. Para ello habrá que invertir la función logit, que viene dada de la siguiente manera

$$g^{-1}(x) = \frac{e^x}{1 + e^x}$$

Finalmente, el modelo logístico consiste en expresar la probabilidad de éxito de la siguiente manera:

$$\pi(x, \beta) = g^{-1}(x' \beta) = \frac{e^{x' \beta}}{1 + e^{x' \beta}}$$

Se define la Odds Ratio (OR), como el cociente de la odds en una y otra población y se representa como $OR = e^{\beta}$. Cuantifica la magnitud de la asociación entre la respuesta y el factor de interés.

3.7. REGRESIÓN MULTINOMIAL

La regresión logística multinomial (Hosmer y Lemeshow, 1989) se utiliza en modelos con variable dependiente de tipo nominal con más de dos categorías (policotómica) y es una extensión multivariante de la regresión logística binaria.

En el modelado de una respuesta policotómica nominal con k categorías, se deben definir $k-1$ logits, cada uno de ellos comparando una determinada categoría de k de respuesta con la categoría de referencia, codificada con un 0. Por tanto, el modelo de regresión logística multinomial plantea un conjunto de $k-1$ ecuaciones de regresiones simultáneas:

$$\text{logit}(Y = s) = \ln \left(\frac{P(Y = s)}{P(Y = 0)} \right) = \beta_{0s} + \beta_{1s}x \quad \text{siendo } s = 1, \dots, k - 1$$

3.8. ANÁLISIS DE SUPERVIVENCIA

Introducción

El análisis de supervivencia consiste en un conjunto de técnicas para analizar el tiempo de seguimiento hasta la ocurrencia de un evento de interés (en nuestro caso muerte). En la práctica, es muy frecuente encontrarse con situaciones que se cuentan con observaciones incompletas de los períodos que transcurren entre el tiempo inicial y el tiempo final. Esto puede darse por censura o por truncamiento, y es precisamente bajo censura o truncamiento que el análisis de supervivencia cobra una mayor importancia. Los mecanismos de censura y truncamiento más frecuentes son los siguientes:

- i. Censura por la derecha: Se presenta cuando hasta la última observación que se le hace al individuo, aún no ha ocurrido el evento que se desea observar. Existen varias razones para que se presente este tipo de censura:
 - Que hasta el momento de finalización del estudio no haya ocurrido el evento, esto ocurriría en el caso de que el período de seguimiento sea finito.
 - Que el individuo haya abandonado el estudio.
 - Que haya ocurrido en el individuo otro evento que imposibilite la ocurrencia del evento que se desea observar.
- ii. Censura por la izquierda: Es poco común en el análisis de supervivencia, se presenta cuando para la primera observación que se realiza sobre el individuo ya ha ocurrido el evento de interés. Este tipo de censura suele confundirse con el truncamiento por la izquierda o la entrada tardía.
- iii. Censura por intervalos: Se presenta cuando solo se sabe que al individuo le ocurre el evento de interés entre un instante t_i y un tiempo t_j .

- iv. Entrada tardía al estudio (truncamiento por la izquierda): Se presenta cuando el individuo comienza a observarse posteriormente al verdadero evento inicial.
- v. Truncamiento por la derecha: Se presenta cuando sólo se incluyen los individuos que presentan el evento de interés.

Para obtener un panorama general de los distintos tipos de censura puede verse el libro de (Andersen et al., 1993) o el de (Klein y Moeschberger, 1997).

Definiciones básicas

Función de Supervivencia

La función de supervivencia se define como la probabilidad de que una persona sobreviva (no le ocurra el evento de interés) al menos hasta el tiempo t . Una definición más formal puede darse de la siguiente manera: sea T una variable aleatoria positiva con función de distribución $F(t)$ y función de densidad $f(t)$. La función de supervivencia $S(t)$ es:

$$S(t) = 1 - F(t) = P[T > t]$$

Funciones de Riesgos

La función de razón de riesgos $\lambda(t)$ se define como el cociente entre la función de densidad y la función de supervivencia:

$$\lambda(t) = \frac{f(t)}{S(t)}$$

Se interpreta como la probabilidad de que a un individuo le ocurra el evento de interés en la siguiente unidad de tiempo Δt dado que ha sobrevivido al tiempo t .

La función de riesgo acumulada $\Lambda(t)$ se define como:

$$\Lambda(t) = \int_0^t \lambda(u) du = -\log S(t)$$

Como habíamos planteado anteriormente, lo que distingue el análisis de supervivencia es la presencia de censura, que en nuestro caso será la de censura por la derecha.

Los datos de supervivencia suelen presentarse de la forma (t_i, δ_i) , donde t_i es el tiempo de observación y, $\delta_i = 0$ si la observación es censurada y $\delta_i = 1$ cuando se observa la ocurrencia del evento de interés.

Estimador de Kaplan-Meier

La presencia de datos censurados o truncados hace que la función de supervivencia no pueda ser obtenida directamente a través de argumentos probabilísticos haciéndose necesario el uso de algunos estimadores. Existen varias formas de estimar la función de supervivencia, entre los más conocidos son los basados en tablas de vida, entre el que se incluye el estimador actuarial y el estimador de Kaplan-Meier, que es más práctico, ya que no es necesario trabajar con períodos de tiempos, sino que los mismos tiempos de observación van contribuyendo a la estimación de la función de supervivencia.

El estimador de Kaplan-Meier (Kaplan-Meier, 1958) es el estimador de la función de supervivencia más utilizado y se define para el caso en que los datos puedan presentar censura por la derecha como:

$$\hat{S}_{KM}(t) = \prod_{t_i \leq t} \frac{r(t_i) - d(t_i)}{r(t_i)}$$

donde $r(t_i)$ y $d(t_i)$ son el número de individuos en riesgo y el número de muertes en el momento t_i .

La varianza del estimador de Kaplan-Meier se obtiene a través de la fórmula de Greenwood (1926):

$$V(\hat{S}_{KM}(t)) = \hat{S}_{KM}^2(t) \sum_{t_i \leq t} \frac{d(t_i)}{r(t_i)[r(t_i) - d(t_i)]}$$

El intervalo de confianza al 95% se obtiene mediante:

$$\hat{S}_{KM}(t) \pm 1.96 ee(\hat{S}_{KM}(t))$$

donde $ee(\hat{S}_{KM}(t))$ es el error estándar de estimación del estimador de Kaplan-Meier.

Comparación de las funciones de supervivencia

La comparación de curvas se efectúa a través de contrastes basados en tablas de contingencia, en nuestro caso, la tabla es de la siguiente forma:

Tabla.1

	Grupo		
Evento	1	0	Total
Muerte	$d_1(t_i)$	$d_0(t_i)$	$d(t_i)$
No Muerte	$r_1(t_i) - d_1(t_i)$	$r_0(t_i) - d_0(t_i)$	$r(t_i) - d(t_i)$
En riesgo	$r_1(t_i)$	$r_0(t_i)$	$r(t_i)$

Se han definido los grupos, como 1 (Controlado) y 0 (No controlado), correspondiendo estos grupos a cada una de las curvas de supervivencia.

Para construir el estadístico de contraste basta con calcular el número esperado de muertes y la varianza estimada del número de muertes para cada uno de los grupos; por ejemplo, para el grupo 1. El número esperado de muertes se calcula de la siguiente manera:

$$\hat{e}_1(t_i) = \frac{r_1(t_i)d(t_i)}{r(t_i)}$$

La varianza estimada de $d_1(t_i)$ está basada en la distribución hipergeométrica y para el grupo 1 está definida como:

$$\hat{V}(d_1(t_i)) = \frac{r_1(t_i)r_0(t_i)(r(t_i) - d(t_i))}{r^2(t_i)(r(t_i) - 1)}$$

Finalmente, el estadístico de contraste se define de la siguiente manera:

$$Q = \frac{\left[\sum_{i=1}^m w_i (d_1(t_i) - \hat{e}_1(t_i)) \right]^2}{\sum_{i=1}^m w_i^2 \hat{V}(d_1(t_i))}$$

Este estadístico se puede aproximar mediante una chi-cuadrado de un grado de libertad si el número de ocurrencias de eventos es grande.

Bajo la hipótesis nula, se asume que las dos funciones de supervivencia son iguales. En esta fórmula m es el número de tiempos de ocurrencia de eventos en ambos grupos y w_i denota los pesos, que toman valores distintos dependiendo del test utilizado. Utilizaremos el más común de los tests, que es el de Mantel y Haenzel (o long-rank). Este test está diseñado para verificar igualdad o diferencia en la función de supervivencia en todos los tiempos. En este test los pesos son iguales a 1, es decir, $w_i = 1$. (Mantel, 1966).

Otro test comúnmente utilizado es el de Peto y Peto (1972). Este test permite verificar igualdad o diferencia de las funciones de supervivencia en los tiempos iniciales. Para una enumeración muy completa de los distintos test, basados en procesos de conteo (Andersen et al, 1993) y (Fleming y Harrington, 1991).

Supervivencia media y mediana

Supervivencia media

La supervivencia media puede ser estimada mediante la siguiente expresión:

$$\hat{\mu} = \int_0^T \hat{S}_{KM}(t) dt$$

donde T es tiempo máximo de seguimiento observado durante el estudio.

Supervivencia mediana

La supervivencia mediana se define como el primer tiempo t que satisface la siguiente condición:

$$\hat{S}_{KM}(t) \leq 0.5$$

Modelo de regresión de Cox

El modelo de regresión de Cox (Cox, 1972) es el modelo de regresión más utilizado para datos de supervivencia en el área médica.

En el modelo de regresión de Cox, el riesgo para el i -ésimo individuo se define mediante la siguiente expresión:

$$\lambda(t; Z_i(t)) = \lambda_0(t) e^{\beta' Z_i(t)}$$

donde $Z_i(t)$ es el vector de covariables para el i -ésimo individuo en el tiempo t .

El modelo de Cox establecido anteriormente se dice que es un modelo semiparamétrico debido a que incluye una parte paramétrica y otra no paramétrica.

- i. La parte paramétrica es $r_i(t) = e^{\beta' Z_i(t)}$, llamada puntaje de riesgo (risk score), y β es el vector de parámetros de la regresión.
- ii. La parte no paramétrica es $\lambda_0(t)$ que es llamada función de riesgo base, es una función arbitraria y no especificada.

El modelo de regresión de Cox se llama también modelo de riesgos proporcionales debido a que el cociente entre el riesgo para dos sujetos con el mismo vector de covariables es constante en el tiempo, es decir:

$$\frac{\lambda(t; Z_i(t))}{\lambda(t; Z_j(t))} = \frac{\lambda_0(t)e^{\beta'Z_i(t)}}{\lambda_0(t)e^{\beta'Z_j(t)}} = \frac{e^{\beta'Z_i(t)}}{e^{\beta'Z_j(t)}}$$

Suponiendo que una muerte ha ocurrido en el tiempo t^* , entonces la verosimilitud de que la muerte le ocurra al individuo i -ésimo y no a otro individuo es:

$$L_i(\beta) = \frac{\lambda_0(t^*)r_i(t^*)}{\sum_j Y_j(t^*)\lambda_0(t^*)r_j(t^*)} = \frac{r_i(t^*)}{\sum_j Y_j(t^*)r_j(t^*)}$$

El producto de los términos de la expresión anterior $L(\beta) = \prod_i L_i(\beta)$ se denomina la verosimilitud parcial y fue introducida por Cox.

La maximización de $\log(L(\beta))$ da una estimación para β sin necesidad de estimar el parámetro de ruido o función de riesgo base $\lambda_0(t)$.

Contrastes de hipótesis para el modelo de Cox.

Una vez que se ha ajustado el modelo Cox, existen tres contrastes de hipótesis para verificar la significación del modelo, estos tests son asintóticamente equivalentes, pero no siempre sucede lo mismo en la práctica. Estos tres contrastes son el test de razón de verosimilitudes, es el que presenta una mayor confiabilidad, el test de Wald, es quizás el más natural debido a que proporciona un contraste por variables en vez de una medida de significación global, y el test de puntuajes (score test).

Modelos de Cox estratificados

Una extensión del modelo de Cox permite obtener la estimación de los modelos para distintos grupos disjuntos o estratos. El modelo obtenido se conoce como modelo de Cox estratificado y está definido para el estrato j -ésimo como:

$$\lambda(t; Z_i(t)) = \lambda_j(t)e^{\beta'Z_i(t)}$$

Este modelo permite obtener la estimación del modelo en presencia de una variable de estratificación sobre la cual se desean obtener funciones de supervivencia por cada uno de los distintos grupos y probablemente poder estudiar la existencia o no de las funciones de supervivencia entre los grupos.

El modelo de Cox estratificado también constituye una de las maneras de corregir el modelo de Cox cuando no se cumple el supuesto de riesgos proporcionales para alguna de las covariables. En este caso suele correrse el modelo estratificando por la covariable que no cumple con el supuesto de riesgo proporcional. Este procedimiento permite corregir el sesgo en la estimación del parámetro que puede presentarse cuando se viola el supuesto de riesgo proporcional. Sin embargo, presenta una desventaja y es que no existe ningún β que permita estimar el efecto de la covariable de estratificación.

Estudio de residuos en el análisis de supervivencia

Una de las ventajas que han surgido del enfoque del análisis de supervivencia es la posibilidad de efectuar análisis de residuos (Andersen et al., 1993, Fleming y Harrington, 1991, Therneau y Grambsch, 2000, Therneau et al., 1990).

Los residuos se pueden utilizar para:

1. Descubrir la forma funcional correcta de un predictor continuo.
2. Identificar los sujetos que están pobremente predichos por el modelo.
3. Identificar los puntos o individuos de influencia.
4. Verificar el supuesto de riesgo proporcional.

Existen cuatro tipos de residuos de interés en el modelo de Cox: los residuos de martingala, los de desvíos (deviance), los de puntaje (score) y los de Schoenfeld.

Residuos de Martingala

Los residuos de martingala se usan para estudiar la forma funcional de una covariable.

Residuos de desvíos (deviance)

Los residuos de desvíos se utilizan para la detección de valores atípicos (outliers).

Residuos de puntajes (scores)

Los residuos de puntajes se utilizan para verificar la influencia individual y para la estimación robusta de la varianza.

Residuos de Schoenfeld

Los residuos de Schoenfeld (Schoenfeld, 1982) son útiles para la verificación del supuesto de riesgo proporcional en el modelo de Cox.

4. RESULTADOS

4.1. Características clínicas de los individuos según el control

En la Figura.1 se explica mediante un diagrama como se definen los pacientes controlados y los no controlados en la consulta implantada en el servicio de Medicina interna. En la Tabla.1 se muestra las características clínicas de los pacientes controlados (que además de la primera consulta asistieron al menos a la segunda consulta) y de los no controlados (que asistieron solamente a la primera consulta). El grupo no controlado contiene: un número mayor de individuos que presentan anemia ($P < 0.05$) y un número menor de pacientes que presentan ninguna o poca limitación en la actividad física ($P < 0.05$). Los resultados de la Tabla.1 se obtuvieron utilizando la función *test.mixto* y la función *test.cualitativa* definida en el Capítulo 3.

Figura.1 Control de los individuos en la consulta especializada

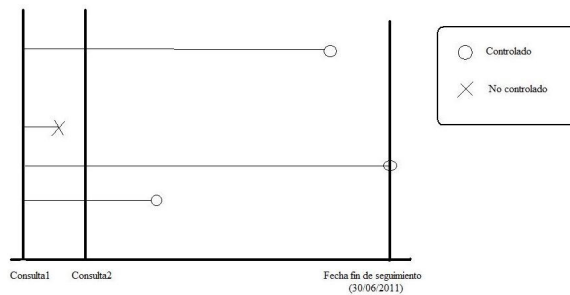


Tabla.1 Características clínicas de los pacientes

	Controlados (n=255)	No controlados (n=22)	P-valor
Edad (años)	79 ± 8	80 ± 6	0.8755
Sexo (Hombre)	117 (46%)	11 (50%)	0.7101
Hipertensión (Si)	212 (83%)	14 (64%)	0.0396
Diabetes (Si)	81 (32%)	10 (45%)	0.1896
EPOC (Si)	75 (29%)	10 (45%)	0.1175
FA (Si)	140 (55%)	15 (68%)	0.2287
Glucemia (mg/dl)	114 ± 38	142 ± 74	0.0661
Hemoglobina (g/dl)	13 ± 2	12 ± 2	0.0803
Sodio (mMol/l)	139 ± 3	138 ± 4	0.2502
GFR (ml/min)	57 ± 21	56 ± 19	0.9315
Hiponatremia (Si)	27 (11%)	4 (18%)	0.2867
Anemia (Si)	43 (17%)	9 (41%)	0.0102
ERC (Si)	148 (58%)	10 (45%)	0.2526
NYHA (1)	180 (71%)	11 (50%)	0.0389

En cada grupo para las variables cuantitativas se representa la media y su desviación típica y para las variables cualitativas las frecuencias y sus porcentajes. La variable NYHA=1 indica ninguna o poca limitación en la actividad física.

4.2. Características clínicas de los individuos según el status

En la Tabla.2 se muestra las características clínicas de los individuos según el status, es decir de los pacientes muertos y de los vivos. El grupo de los individuos que fallecieron presenta una mayor edad que los que viven ($P < 0.05$), un menor nivel de glucosa, hemoglobina y sodio que los que viven ($P < 0.05$) un mayor porcentaje de individuos que presentan anemia ($P < 0.05$) y un menor porcentaje de pacientes que presentan ninguna o poca limitación en la actividad física ($P < 0.001$).

En las Figura.2, Figura.3, Figura.4 y Figura.5 se representan los diagramas de cajas y las estimaciones de las densidades de las variables edad, glucemia, hemoglobina y sodio respectivamente según la variable categórica *status* (muerte=1, vivo=0).

En la Figura.6 se representa los diagramas de barra de las frecuencias absolutas de anemia y NYHA según la variable categórica *status*.

Los resultados de la Tabla.2 y todas las figuras de esta sección 4.2 se obtuvieron utilizando las funciones *test.mixto* y *test.cualitativa* definidas en el Capítulo3.

Tabla.2 Características clínicas de los pacientes según el status

	Grupo Muerto (n=105)	Grupo Vivo (n=172)	P-valor
Edad (años)	81 ± 7	78 ± 8	0.0001
Sexo (Hombre)	52 (49%)	76 (44%)	0.3873
Hipertensión (Si)	88 (84%)	138 (80%)	0.4561
Diabetes (Si)	29 (28%)	62 (36%)	0.1896
EPOC (Si)	39 (37%)	46 (27%)	0.0687
FA (Si)	62 (59%)	93 (54%)	0.4181
Glucemia (mg/dl)	109 ± 31	121 ± 48	0.0426
Hemoglobina (g/dl)	12 ± 2	13 ± 2	0.0012
Sodio (mMol/l)	138 ± 3	139 ± 3	0.0030
GFR (ml/min)	55 ± 20	58 ± 22	0.2294
Hiponatremia (Si)	14 (13%)	17 (10%)	0.3770
Anemia (Si)	28 (27%)	24 (14%)	0.0086
ERC (Si)	63 (60%)	95 (55%)	0.4368
NYHA (I)	52 (50%)	139 (81%)	0.0000

En cada grupo para las variables cuantitativas se representa la media y su desviación típica y para las variables cualitativas las frecuencias y sus porcentajes. La variable NYHA=1 indica ninguna o poca limitación en la actividad física.

Figura.2 Diagrama de cajas y las diferentes densidades de la Edad en ambos grupos

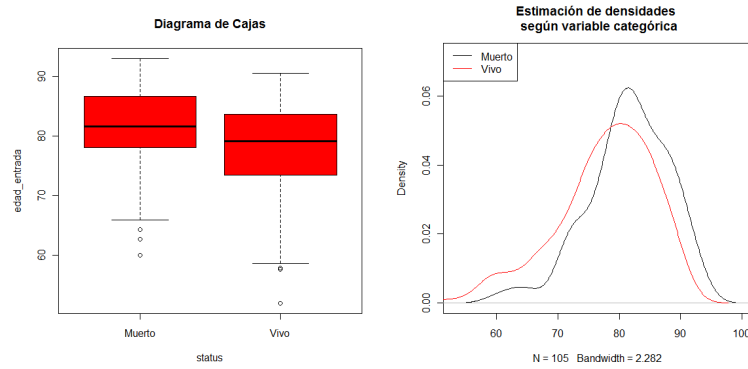


Figura.3 Diagrama de cajas y las diferentes densidades de Glucemia en ambos grupos

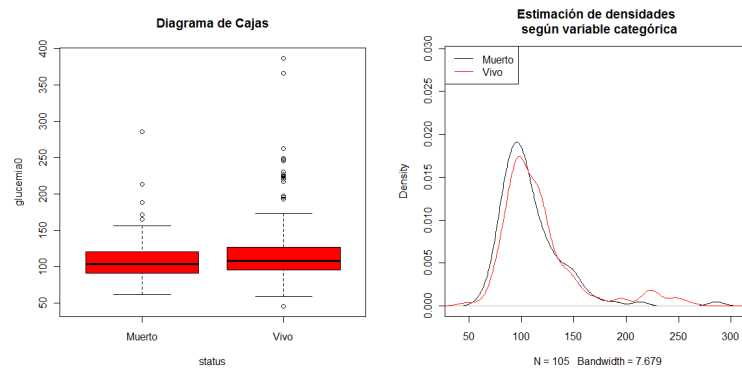


Figura.4 Diagrama de cajas y las diferentes densidades de Hemoglobina en ambos grupos

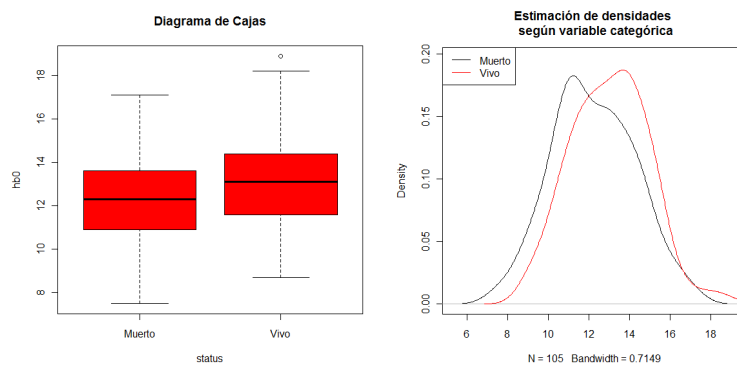


Figura.5 Diagrama de cajas y las diferentes densidades de Sodio en ambos grupos

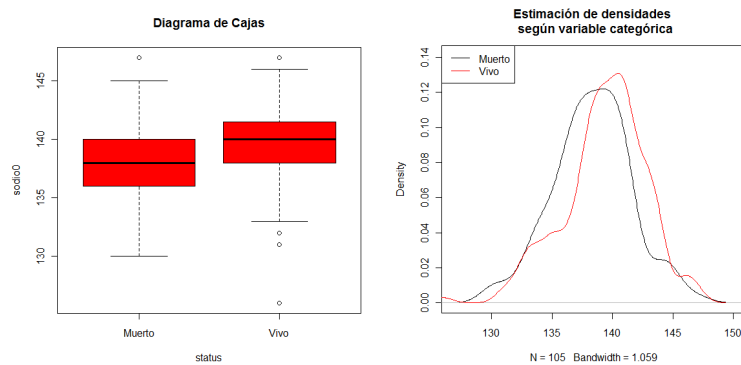
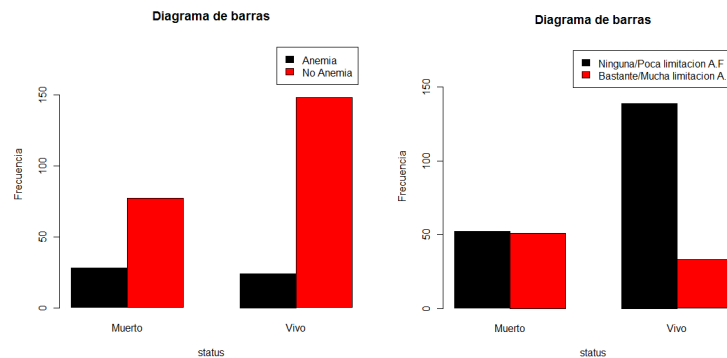


Figura.6 Diagrama de barras de las frecuencias absolutas de Anemia y NYHA en ambos grupos



4.3. Características clínicas de los individuos controlados según los motivos de finalización del seguimiento

En la Figura.7 se explica mediante un diagrama como se definen los motivos de finalización del seguimiento en la consulta especializada. Se diferencian cuatro motivos:

- Seguido: El individuo se controló hasta la fecha fin de seguimiento (30/06/2011).
- Perdida: El individuo dejó de asistir a la consulta por causas ajenas a la HF.
- Muerte: El individuo dejó de asistir a la consulta porque falleció.
- Alta: El individuo dejó de asistir a la consulta porque se derivó a atención primaria.

En la Tabla.3 se muestra las características clínicas según los motivos de finalización del seguimiento. Los individuos que son seguidos y los que les dan alta presentan una mayor cantidad de sodio que los que se son perdidos. El grupo de seguidos y de altas tienen un mayor porcentaje de individuos con ninguna o poca limitación de la actividad física que el grupo muerte.

En la Figura.8 y la Figura.9 se representa el diagrama de cajas junto con la estimación de la densidad de la variable sodio y el diagrama de barras de las frecuencias absolutas de la variable NYHA respectivamente, según la variable categórica *motivo*.

Los resultados de la Tabla.3 y todas las figuras de esta sección 4.3 se obtuvieron utilizando las funciones *test.mixto* y *test.cualitativa* definidas en el Capítulo3.

Figura.7 Motivos finalización del seguimiento

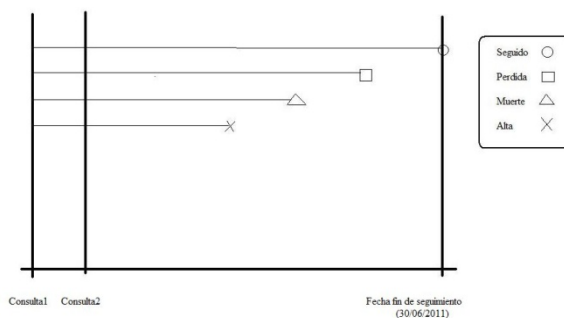


Tabla.3 Características clínicas según los motivos de finalización de seguimiento

	Seguido (n=96)	Perdida (n=71)	Muerte (n=34)	Alta (n=54)	P-valor
Edad (años)	78 ± 7	80 ± 9	80 ± 7	79 ± 7	0.1
Sexo (Hombre)	39 (41%)	35 (49%)	17 (50%)	26 (48%)	0.6268
Hipertensión (Si)	81 (84%)	59 (83%)	29 (85%)	43 (80%)	0.8752
Diabetes (Si)	31 (32%)	20 (28%)	10 (29%)	20 (37%)	0.7494
EPOC (Si)	27 (28%)	17 (24%)	14 (41%)	17 (31%)	0.3237
FA (Si)	55 (57%)	37 (52%)	23 (68%)	25 (46%)	0.2318
Glucemia (mg/dl)	117 ± 42	111 ± 36	113 ± 40	112 ± 32	0.593
Hemoglobina (g/dl)	13 ± 2	13 ± 2	12 ± 2	13 ± 2	0.6075
Sodio (mMol/l)	140 ± 3	138 ± 3	139 ± 3	139 ± 3	0.0016
GFR (ml/min)	56 ± 19	58 ± 24	57 ± 18	59 ± 23	0.7668
Hiponatremia (Si)	6 (6%)	13 (18%)	4 (12%)	4 (7%)	0.0717
Anemia (Si)	11 (11%)	14 (20%)	7 (21%)	11 (20%)	0.3584
ERC (Si)	55 (57%)	43 (61%)	19 (56%)	31 (57%)	0.9635
NYHA (1)	73 (76%)	46 (66%)	18 (55%)	43 (80%)	0.0376

En cada grupo para las variables cuantitativas se representa la media y su desviación típica y para las variables cualitativas las frecuencias y sus porcentajes. La variable NYHA=1 indica ninguna o poca limitación en la actividad física.

Figura.8 Diagrama de cajas y las diferentes densidades de Sodio en ambos motivos

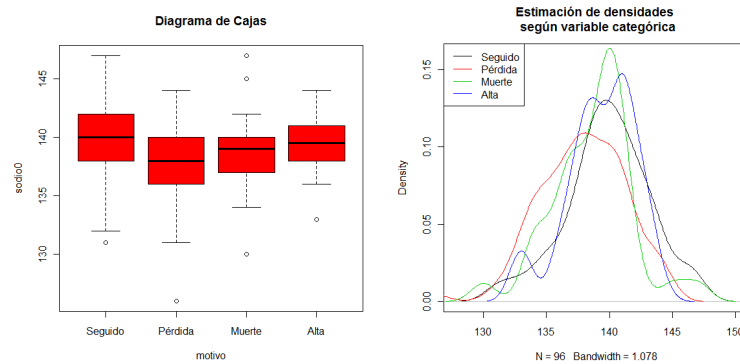
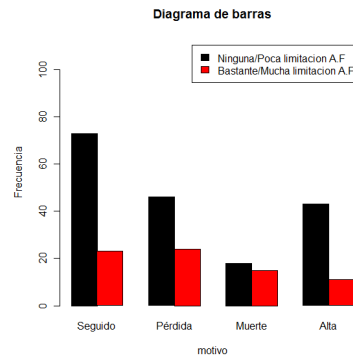


Figura.9 Diagrama de barras de las frecuencias absolutas de la NYHA en ambos motivos



4.4. Repercusión del seguimiento

En la Tabla.4 se comparan las variables basales con las finales en pacientes que se controlaron hasta la fecha fin de seguimiento (30/06/2011). Estos pacientes son los definidos como los 96 seguidos de la sección 4.3.

De la Tabla.4 se observa que el porcentaje de enfermos renales crónicos aumenta significativamente y que el estado funcional de esos pacientes que se controlaron hasta la fecha fin de seguimiento mejora significativamente.

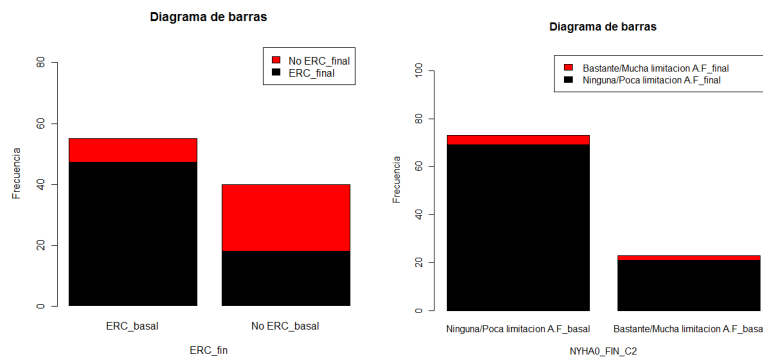
Los resultados de la Tabla.4 y las figuras de esta sección 4.4 se obtuvieron utilizando las funciones: *test.cuantitativa* y *test.cualitativa* del Capítulo3.

Tabla.4 Comparación basal y final

	Basal	Final	P-valor
Glucemia (mg/dl)	117 ± 42	114 ± 34	0.3261
Hemoglobina (g/dl)	13 ± 2	13 ± 2	0.6462
Sodio (mMol/l)	140 ± 3	141 ± 3	0.0561
GFR (ml/min)	56 ± 19	52 ± 17	0.0648
Hiponatremia (Si)	6 (6%)	2 (2%)	0.1336
Anemia (Si)	11 (11%)	9 (9%)	0.7237
ERC (Si)	55 (57%)	65 (68%)	0.0499
NYHA (I)	73 (76%)	90 (94%)	0.0007

En cada grupo para las variables cuantitativas se representa la media y su desviación típica y para las variables cualitativas las frecuencias y sus porcentajes. La variable NYHA=1 indica ninguna o poca limitación en la actividad física.

Figura.10 Diagrama de barras de las frecuencias absolutas basal y final de ERC y NYHA



El diagrama de la izquierda de la Figura.10, representa el cambio que se produce en la variable ERC. De los 55 enfermos renales crónicos que había al principio, 47 siguieron enfermos y 8 no siguieron enfermos al final, mientras que de los 40 que no eran enfermos renales crónicos al principio, 22 siguen sin serlo y 18 pasan a serlo al final. Por lo que el porcentaje de enfermos renales crónicos aumenta significativamente.

El diagrama de la derecha de la Figura.10, representa el cambio que se produce en la variable NYHA. De los 73 individuos que había al principio con ninguna o poca limitación en la actividad física, 69 siguieron igual y 4 pasaron a tener bastante o mucha limitación al final, mientras que de los 23 individuos que había al principio con bastante o mucha limitación en la actividad física, 2 siguieron teniéndola y 21 pasaron a tener ninguna o poca bastante limitación. Esto indica que los pacientes que se controlaron hasta la fecha fin de estudio, su estado funcional mejora significativamente por lo que el seguimiento mejora la calidad de vida.

4.5. Factores pronóstico de la mortalidad

En la Tabla.5 se muestra los coeficientes estimados del modelo, el error estándar de estas estimaciones, el p-valor y el riesgo relativo con sus intervalos de confianza al 95%. En la Figura.11 representamos los efectos que tienen las variables del modelo en la mortalidad, para ello utilizamos la librería effects (John Fox, 2003).

Tabla.5 Estimación de los coeficientes del modelo

	Coeficientes	SE	P-valor	Riesgo Relativo (RR)	I.C. 95% del R.R
Constante	11.45	6.02	0.0574		
Edad	0.09	0.02	0.0001	1.09	(1.04, 1.14)
Sexo (Mujer)	-0.55	0.30	0.0653	0.58	(0.32, 1.03)
Hipertension (Si)	0.65	0.39	0.0925	1.91	(0.91, 4.18)
Glucemia	-0.01	0.00	0.0647	0.99	(0.98, 1.00)
Hemoglobina	-0.14	0.07	0.0471	0.87	(0.75, 0.99)
Sodio	-0.12	0.04	0.0062	0.89	(0.81, 0.96)
NYHA (3)	1.40	0.30	0.0000	4.05	(2.25, 7.42)

En esta tabla se muestra los coeficientes estimados del modelo, el error estándar (SE), el P-valor y el riesgo relativo con sus intervalos de confianza al 95%. La variable NYHA=3 indica bastante o mucha limitación en la actividad física.

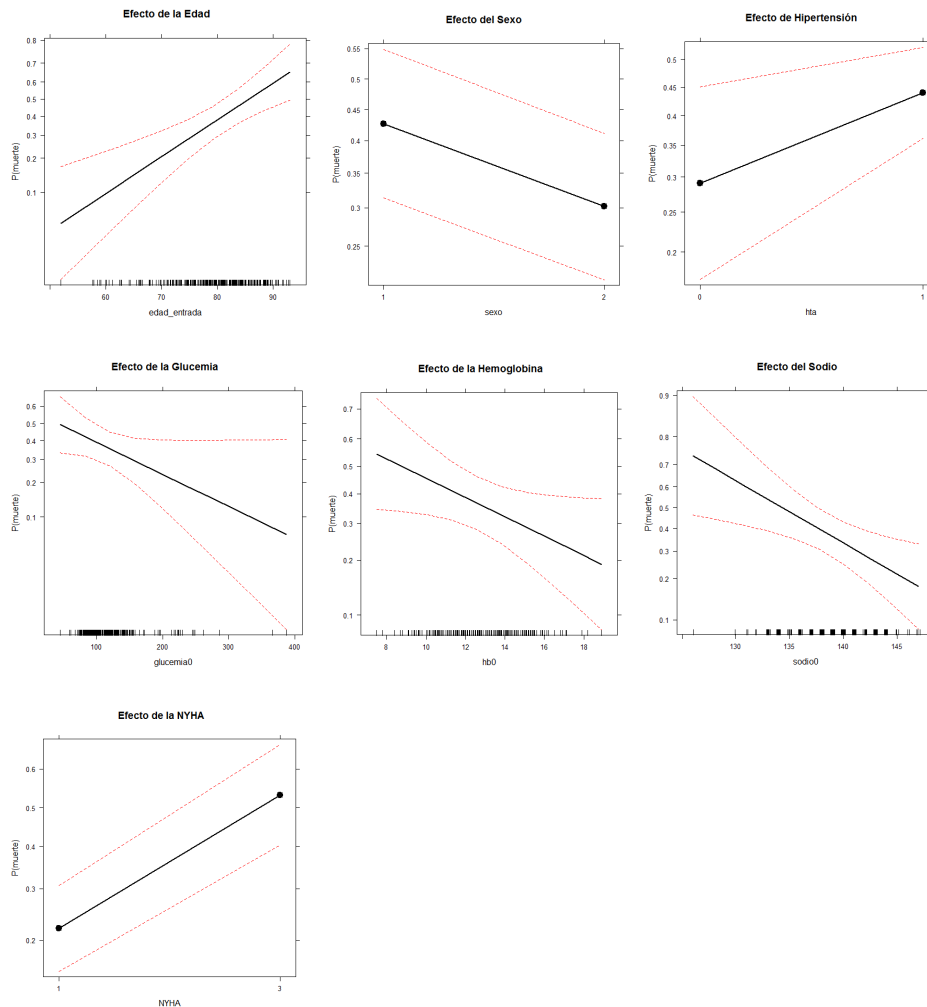
La estimación del RR respecto a la edad es 1.09, con un I.C. al 95% (1.04, 1.14). Por cada año que aumenta la edad se multiplica por 1.09 la probabilidad de morir. Esta relación es significativa, pues el intervalo de confianza no contiene al 1 y ($P < 0.05$).

La estimación del RR por cada unidad que aumenta el nivel de hemoglobina es 0.87, es decir, por cada unidad menos en el nivel de hemoglobina se multiplica por 1.15 la probabilidad de morir. El intervalo de confianza para este RR es (0.75, 0.99) como podemos observar no contiene el valor 1 y por tanto refleja una relación significativa. Este hecho lo confirma ($P < 0.05$).

La estimación del RR por cada unidad que aumenta el nivel de sodio es 0.89, es decir, por cada unidad menos en el nivel de sodio se multiplica por 1.13 la probabilidad de morir. El intervalo de confianza para este RR es (0.81, 0.96) como podemos observar no contiene el valor 1 y por tanto refleja una relación significativa. Este hecho lo confirma ($P < 0.05$).

El RR de NYHA, bastante o mucha limitación en la actividad física respecto con ninguna o poca limitación en la actividad física es 4.048, por lo que los individuos que tengan bastante o mucha limitación tienen 4 veces más riesgo de morir que los que tienen poca o ninguna limitación. El intervalo de confianza es (2.25, 7.42). Este intervalo no contiene el valor 1 por lo que la relación es significativa hecho que también indica ($P < 0.01$).

Figura.11 Efectos de las variables del modelo



4.6. Factores pronóstico de los motivos de finalización del seguimiento

Los motivos de finalización de seguimiento son: Seguido (1), Pérdida (2), Muerte (3) y Alta (4). Son los que hemos visto en la sección 4.3. Definimos 3 logits, cada uno de ellos comparando un determinado motivo de la respuesta con el motivo de referencia, Seguido (1). En la Tabla.6 se muestra los coeficientes estimados del modelo, el error estándar de estas estimaciones y el riesgo relativo con sus intervalos de confianza al 95% para cada uno de los 3 logits. En la Tabla.7 se muestra las probabilidades estimadas en cada uno de los motivos para la variable NYHA.

En la Figura.12, Figura.13 y en la Figura.14 se representa las probabilidades estimadas de cada motivo para las variables edad, sodio y NYHA.

La representación de las figuras y los resultados de las tablas se obtienen, utilizando la librería effects, mencionada en la sección 4.5 y la función multinom de la librería. (Fox y Jangman, 2009).

Tabla.6 Estimación de los coeficientes del modelo de regresión logística multinomial

	Coefficientes	SE	Riesgo Relativo (RR)	IC 95% del R.R
2:Constante	22.60	4.31		
3:Constante	11.19	1.06		
4:Constante	6.33	2.38		
2:Edad_entrada	0.06	0.02	1.06	(1.02, 1.11)
3:Edad_entrada	0.05	0.03	1.05	(0.99, 1.11)
4:Edad_entrada	0.02	0.02	1.02	(0.98, 1.07)
2:Sexo (Mujer)	-0.51	0.32	0.60	(0.32, 1.12)
3:Sexo (Mujer)	-0.63	0.42	0.53	(0.23, 1.22)
4:Sexo (Mujer)	-0.36	0.35	0.69	(0.35, 1.39)
2:Sodio	-0.20	0.03	0.82	(0.77, 0.88)
3:Sodio	-0.11	0.02	0.89	(0.86, 0.92)
4:Sodio	-0.06	0.02	0.94	(0.90, 0.98)
2:NYHA (3)	0.49	0.34	1.63	(0.84, 3.16)
3:NYHA (3)	0.88	0.43	2.40	(1.03, 5.60)
4:NYHA (3)	-0.26	0.42	0.77	(0.34, 1.75)

En esta tabla se muestra los coeficientes estimados del modelo, el error estándar (SE), el P-valor y el riesgo relativo con sus intervalos de confianza al 95% para los 3 logits, siendo el motivo de referencia, Perdida(1). La variable NYHA=3 indica bastante o mucha limitación en la actividad física.

La estimación del RR respecto a la edad, cuando el motivo es Perdida, es 1.06, con un I.C. al 95% (1.02, 1.11). Por cada año que aumenta la edad se multiplica por 1.06 la probabilidad de ser Perdida respecto de ser Seguido.

La estimación del RR por cada unidad que aumenta el nivel de sodio, cuando el motivo es Perdida, es 0.82, es decir, por cada unidad menos en el nivel de sodio se multiplica por 1.22 la probabilidad de ser Perdida respecto de ser Seguido. El intervalo de confianza para este RR es (0.77, 0.88) como podemos observar no contiene el valor 1 y por tanto refleja una relación significativa.

La estimación del RR por cada unidad que aumenta el nivel de sodio, cuando el motivo es Muerte, es 0.89, es decir, por cada unidad menos en el nivel de sodio se multiplica por 1.12 la probabilidad de ser Muerte respecto de ser Seguido. El intervalo de confianza para este RR es (0.86, 0.92).

La estimación del RR por cada unidad que aumenta el nivel de sodio, cuando el motivo es Alta, es 0.94, es decir, por cada unidad menos en el nivel de sodio se multiplica por 1.06 la probabilidad de ser Alta respecto de ser Seguido. El intervalo de confianza para este RR es (0.90, 0.98).

La estimación de RR de NYHA, es 2.40, por lo que los individuos que tengan bastante o mucha limitación, multiplica por 2.40 la probabilidad de ser Muerte respecto de ser Seguido. El intervalo de confianza es (1.03, 5.60). Este intervalo no contiene el valor 1 por lo que la relación es significativa

Figura.13 Efecto de la variable Edad en el motivo de finalización del seguimiento

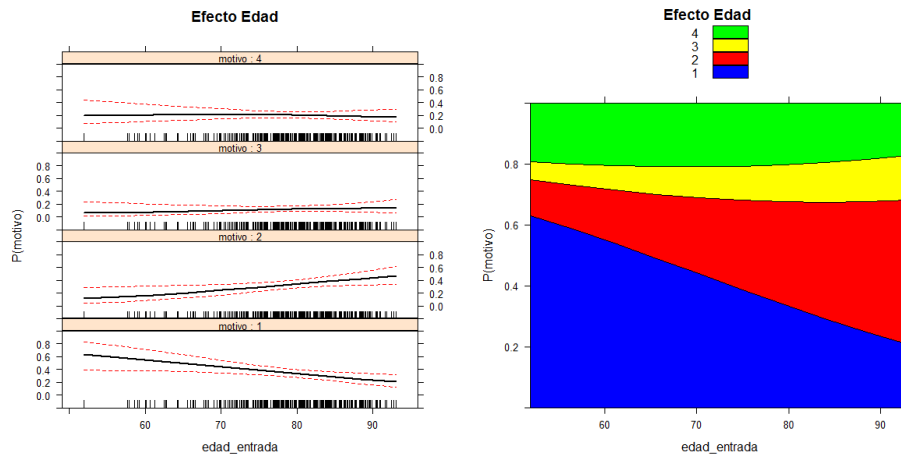


Figura.14 Efecto de la variable Sodio en el motivo de finalización del seguimiento

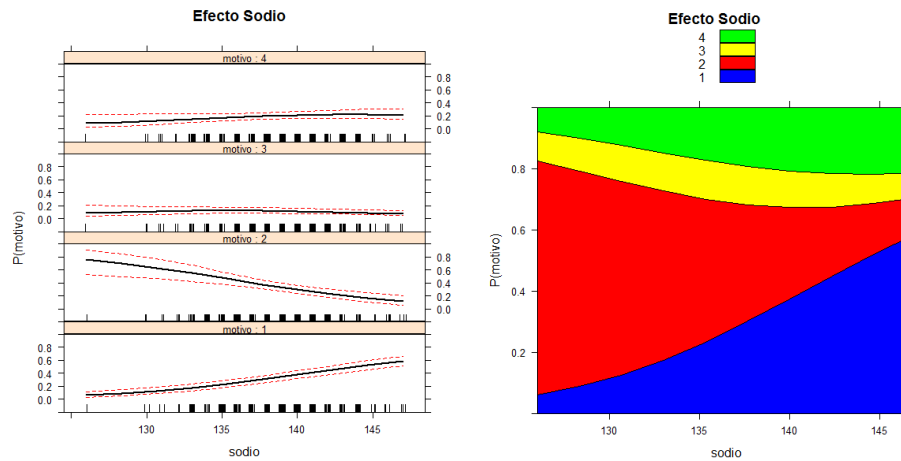


Figura.15 Efecto de la variable NYHA en el motivo de finalización del seguimiento

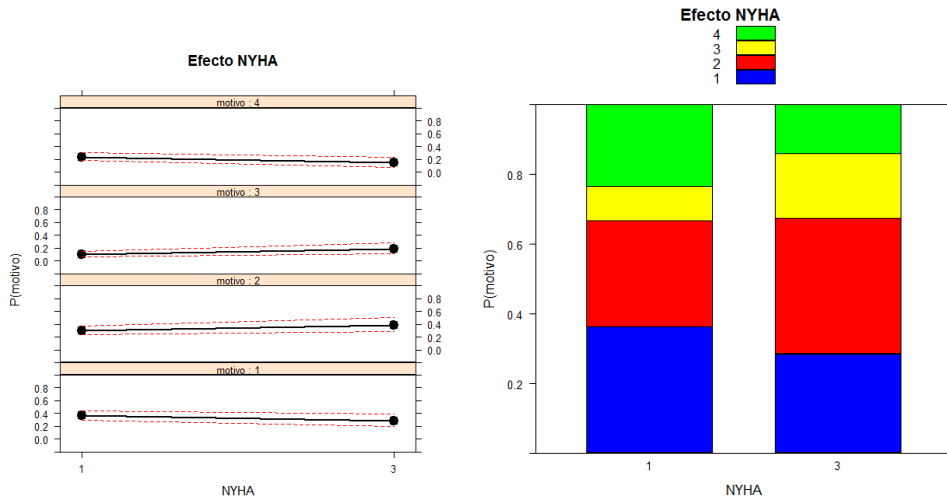


Tabla.7 Probabilidades estimadas en cada uno de los motivos de finalización del seguimiento para la variable NYHA

	Ninguna/Poca Limitación en la Actividad Física	Bastante/Mucha Limitación en la Actividad Física
P(Motivo=Seguido)	0.363	0.284
P(Motivo=Perdida)	0.305	0.389
P(Motivo=Muerte)	0.099	0.186
P(Motivo=Alta)	0.233	0.141

4.7. Valoración de la supervivencia

En la Figura.16 se representa la estimación de la función de Supervivencia y cómo podemos observar, la probabilidad de supervivencia disminuye a medida que pasa el tiempo. En la Figura.17 se representa la supervivencia de una cohorte de pacientes con insuficiencia cardíaca controlados en una consulta especializada.

Figura.16 Estimación de la función de Supervivencia utilizando el método de Kaplan-Meier

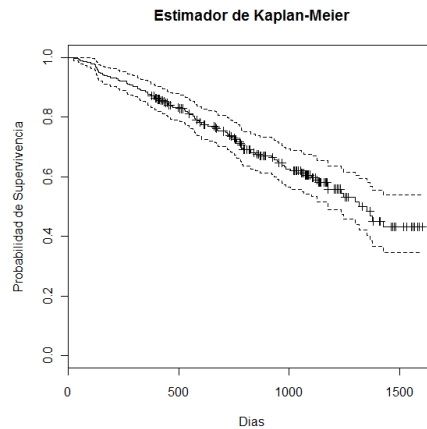
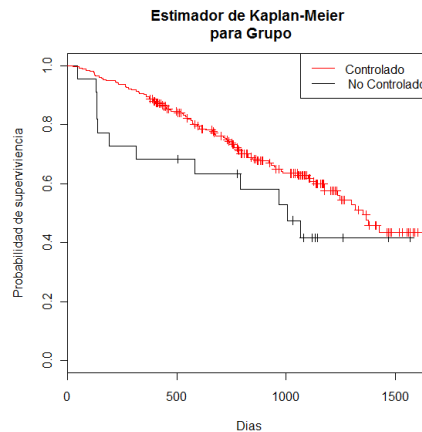


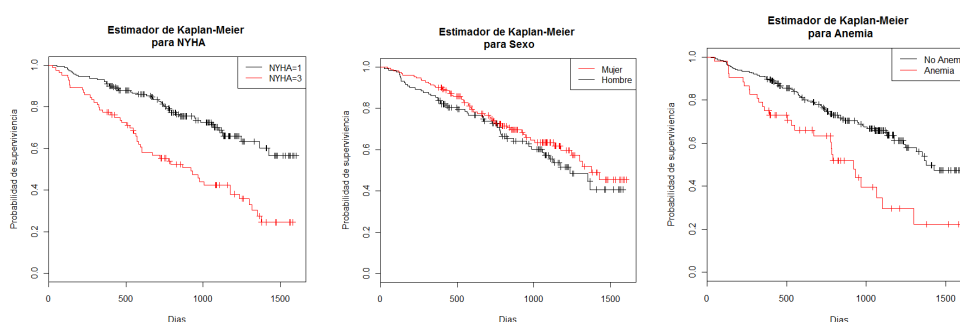
Figura.17 Comparación de funciones de Supervivencia para controlados y no controlados



Aparentemente ambas funciones de supervivencia parecen distintas, para verificar igualdad o diferencia en la función de supervivencia en todos los tiempos utilizamos el test de Mantel y Haenzel (o long-rank). Como $p=0.118 > 0.05$, no hay evidencias significativas para rechazar la hipótesis nula de igualdad de funciones de supervivencia (para un nivel de significación del 5%).

Como el estimador de Kaplan-Meier es un estimador no paramétrico univariante es interesante también verificar igualdad o diferencia en la función de supervivencia para las variables NYHA, sexo, y anemia.

Figura.18 Comparación de funciones de Supervivencia para las variables NYHA, sexo y anemia



Para las variables NYHA, y anemia hay evidencias significativas para rechazar la hipótesis nula de igualdad de funciones de supervivencia (para un nivel de significación del 5%). Tener bastante o mucha limitación en la actividad física afecta negativamente a la supervivencia. Tener anemia afecta negativamente a la supervivencia.

Mientras que para la variable sexo no hay evidencias estadísticamente significativas para rechazar igualdad de funciones de supervivencia para hombres y mujeres.

La regresión de Cox modeliza la función de riesgo, y es una generalización del análisis de Kaplan-Meier. Mientras que con este último método podemos realizar comparaciones entre grupos diferentes definidos por un factor simple, la regresión de Cox ofrece la posibilidad de introducir en el modelo un conjunto de covariables, categóricas o continuas. Así, la regresión de Cox, consiste en obtener una función lineal de las variables independientes que permita estimar, en función del tiempo, la probabilidad de que ocurra el evento de interés (muerte).

En la Tabla.8 se muestra los coeficientes estimados del modelo, el error estándar de estas estimaciones, el p-valor y el riesgo relativo con sus intervalos de confianza al 95%.

Tabla.8 Estimación de los coeficientes del modelo de regresión de Cox

	Coefficientes	SE	P-valor	Riesgo Relativo (RR)	I.C. 95% del R.R
Edad	0.08	0.02	0.0000	1.08	(1.05, 1.12)
Sexo (Mujer)	-0.59	0.20	0.0036	0.55	(0.37, 0.82)
Hemoglobina	-0.14	0.05	0.0088	0.87	(0.78, 0.97)
Sodio	-0.06	0.03	0.0481	0.94	(0.88, 0.99)
NYHA (3)	0.74	0.20	0.0003	2.09	(1.41, 3.11)

En esta tabla se muestra los coeficientes estimados del modelo, el error estándar (SE), el P-valor y el riesgo relativo con sus intervalos de confianza al 95%. La variable NYHA=3 indica bastante o mucha limitación en la actividad física.

La estimación del RR respecto a la edad es 1.08, con un I.C. al 95% (1.05, 1.12). Por cada año que aumenta la edad se multiplica por 1.08 la probabilidad de morir. Esta relación es significativa, pues el intervalo de confianza no contiene al 1 y ($P < 0.01$).

El RR de mujeres respecto a hombres es 0.55, por lo que mujeres tienen menos riesgo de morir que los hombres (los hombres tienen 1.81 veces más riesgo que las mujeres). El intervalo de confianza es (0.37, 0.82). Este intervalo no contiene el valor 1 por lo que la relación es significativa hecho que también indica ($P < 0.01$).

La estimación del RR por cada unidad que aumenta el nivel de hemoglobina es 0.87, es decir, por cada unidad menos en el nivel de hemoglobina se multiplica por 1.15 la probabilidad de morir. El intervalo de confianza para este RR es (0.78, 0.97) como podemos observar no contiene el valor 1 y por tanto refleja una relación significativa. Este hecho lo confirma ($P < 0.01$).

La estimación del RR por cada unidad que aumenta el nivel de sodio es 0.94, es decir, por cada unidad menos en el nivel de sodio se multiplica por 1.07 la probabilidad de morir. El intervalo de confianza para este RR es (0.88, 0.99) como podemos observar no contiene el valor 1 y por tanto refleja una relación significativa. Este hecho lo confirma ($P < 0.05$).

El RR de NYHA, bastante o mucha limitación en la actividad física respecto con ninguna o poca limitación en la actividad física es 2.09, por lo que los individuos que tengan bastante o mucha limitación tienen 2 veces más riesgo de morir que los que tienen poca o ninguna limitación. El intervalo de confianza es (1.41, 3.11). Este intervalo no contiene el valor 1 por lo que la relación es significativa hecho que también indica ($P < 0.01$).

Podemos concluir que hay influencia de la edad, sexo, hemoglobina, sodio y NYHA en la supervivencia tras ser diagnosticado esta enfermedad (Insuficiencia Cardíaca). El menor nivel de hemoglobina y sodio, una mayor edad, ser hombre y tener bastante o mucha limitación en la actividad física afectan negativamente al tiempo de supervivencia.

Los p-valores para el test de razón de verosimilitud, test de Wald y test de los puntajes son menores que 0.05, con lo cual el modelo es significativo.

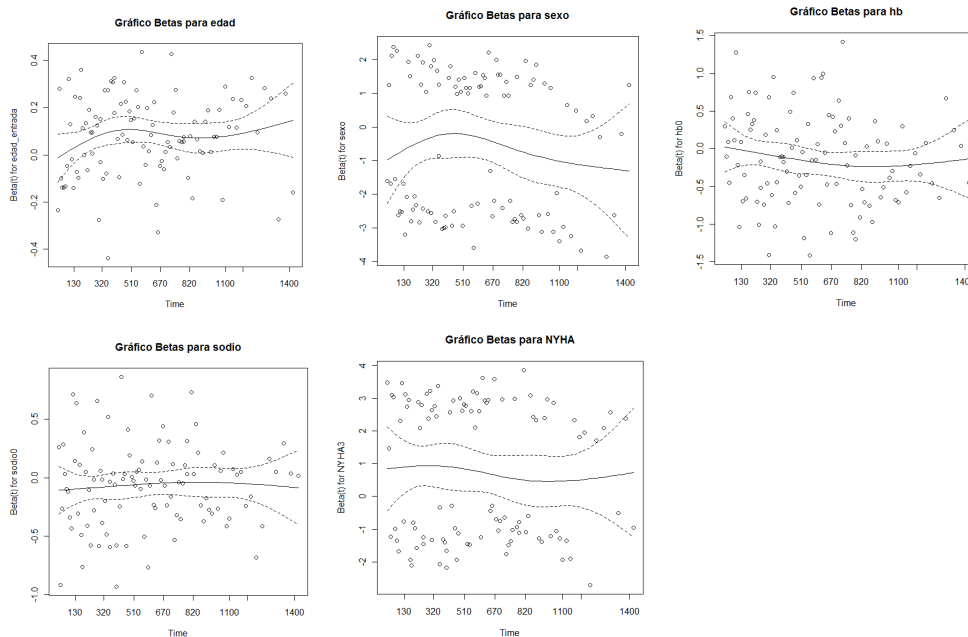
En la Tabla.9 se representa la verificación del supuesto de riesgos proporcionales del modelo de Cox.

Tabla.9 Verificación de los supuestos del modelo de Cox

	<i>rho</i>	<i>chisq</i>	<i>P</i>
Edad_entrada	0.1058	1.242	0.265
Sexo	-0.0909	0.840	0.359
Hemoglobina	-0.1124	1.581	0.209
Sodio	0.0431	0.222	0.637
NYHA	-0.0719	0.536	0.464
Global	NA	4.158	0.527

De donde se concluye de que no existe evidencia significativa al 5% de que se viole el supuesto de riesgos proporcionales, ni desde el punto de vista global, ni para cada covariable.

Figura.19 Gráficos de los betas para cada una de las covariables



En la Figura.20 se representa los residuos tipo deviance, en la Figura.21 se representa los gráficos de influencia sobre la estimación de cada coeficiente y en la Figura.22 se representa la forma funcional de las variables continuas.

En la Figura.20 no existe ningún individuo que esté influenciando en el ajuste del modelo. En la Figura.21 no existe ningún individuo que esté influenciando sobre la estimación del coeficiente correspondiente a las variables edad, sexo, hemoglobina, sodio y NYHA. Y en la Figura.22 la forma funcional de las variables edad, hemoglobina y sodio parece ser la adecuada.

Figura.20 Residuos tipo deviance

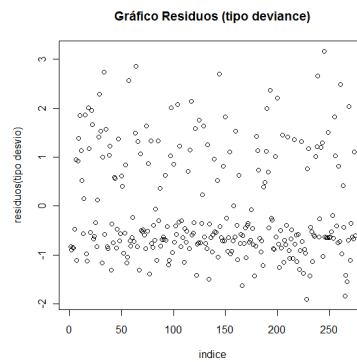


Figura.21 Gráficos de influencia sobre la estimación de cada coeficiente

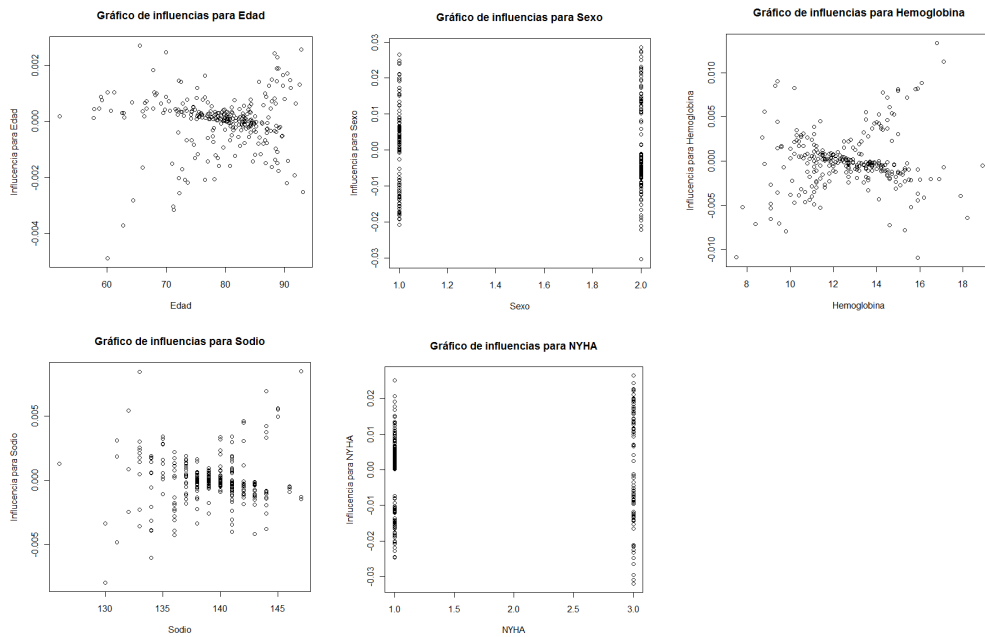
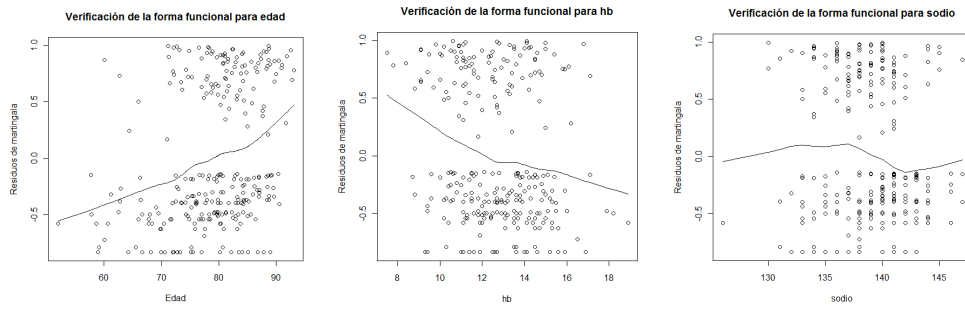
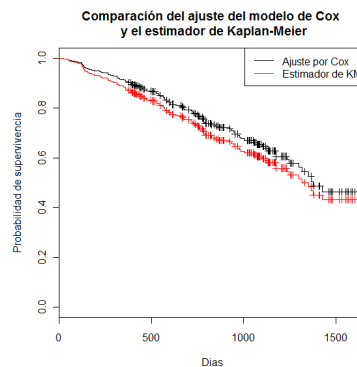


Figura.22 Forma funcional de las variables continuas



En la Figura.23 comparamos el ajuste del modelo de Cox con el estimador de Kaplan-Meier, y podemos ver que la función de supervivencia ajustada por el modelo de Cox es sistemáticamente superior a la función de supervivencia estimada por el método de Kaplan-Meier.

Figura.23 Comparación del ajuste del modelo de Cox con el estimador de Kaplan-Meier



Dependiendo de los propósitos de la investigación o el estudio pudiera ser más adecuado un modelo paramétrico o un modelo de Cox.

Si lo que se pretende es comparar riesgos entre distintos niveles de las covariables, que suele ser el interés de los estudios médicos, probablemente sea más adecuado utilizar un modelo de Cox.

Si el interés está basado en obtener informaciones asociadas con parámetros como medias, varianzas, entre otras, como suele ser el interés en el área de la industria, probablemente se recomiende el uso de un modelo paramétrico.

Nardi y Schemper (2003) plantean una interesante discusión que probablemente ayude a la elección entre un modelo de Cox y un modelo paramétrico.

Se utiliza la librería (survival), para la representación de las figuras que aparecen en esta sección y para calcular los resultados de la Tabla.8 y de la Tabla.9.

5. CONCLUSIONES

El grupo de no controlados presenta un número mayor de individuos que tienen anemia y un número menor de pacientes que tienen ninguna o poca limitación en la actividad física.

El grupo de los individuos que fallecieron presenta una mayor edad que los que viven ($P < 0.05$), un menor nivel de glucosa, hemoglobina y sodio que los que viven ($P < 0.05$) un mayor porcentaje de individuos que presentan anemia ($P < 0.05$) y un menor porcentaje de pacientes que presentan ninguna o poca limitación en la actividad física ($P < 0.001$).

Dentro del grupo de los controlados, los individuos que son seguidos hasta la fecha fin de seguimiento y los que les dan el alta presentan un mayor nivel de sodio que los que se son perdidos. Los individuos que son seguidos hasta la fecha fin de seguimiento y los que les dan el alta, tienen un mayor porcentaje de individuos con ninguna o poca limitación de la actividad física que el grupo muerte.

El estado funcional de los pacientes que fueron seguidos hasta la fecha fin de seguimiento (30/06/2011) mejora significativamente, por lo que el control hasta la fecha fin de seguimiento de los pacientes en la consulta especializada implantada en el servicio de medicina interna, mejora la calidad de vida de estos.

Hay influencia de la edad, hemoglobina, sodio y NYHA en la mortalidad. Un menor nivel de hemoglobina y sodio, una mayor edad, y tener bastante o mucha limitación en la actividad física incrementan el riesgo de morir.

Hay influencia de la edad, sodio y NYHA en el motivo de finalización de seguimiento. Un menor nivel de sodio y una mayor edad incrementan el riesgo de ser perdida respecto de ser seguido. Un menor nivel de sodio y tener bastante o mucha limitación en la actividad física incrementan el riesgo de ser muerte respecto de ser seguido. Un menor nivel de sodio incrementa el riesgo de ser alta respecto de ser seguido.

No hay diferencias significativas en la supervivencia de pacientes con insuficiencia cardíaca (controlados y no controlados) en la consulta especializada puesta en marcha por el servicio de Medicina Interna.

Hay influencia de la edad, sexo, hemoglobina, sodio y NYHA en la supervivencia tras ser diagnosticado esta enfermedad (HF). El menor nivel de hemoglobina y sodio, una mayor edad, ser hombre y tener bastante o mucha limitación en la actividad física afectan negativamente al tiempo de supervivencia.

Valoración de las prácticas

El máster en técnicas estadísticas posee una gran carga teórica, por lo que creo, que un alumno que lo cursa acaba con unos grandes conocimientos de distintas metodologías estadísticas.

La realización de unas prácticas, concretamente en la Unidad de Epidemiología Clínica del CHUS, me parecía que podría ser el complemento ideal a mi formación académica.

Durante la estancia en esta Unidad, además de perfeccionar las técnicas vistas en el máster, he podido aplicar nuevas técnicas estadísticas como por ejemplo: contrastes en tablas de contingencia de datos relacionados (Test de McNemar, Test de Bowker y Test de Stuart y Maxwell), regresión multinomial, ...etc. También trabajé con bases de datos que se tuvo que hacer una depuración previa, antes de realizar un análisis de los datos.

En las prácticas obtuve una visión más general a la hora de realizar un estudio. Es muy importante saber planificar las tareas que se realizarán y tener claro el objetivo final de dicho estudio. Es tan importante saber resolver los problemas, realizar el análisis estadístico, como saber qué problemas resolver, la planificación del estudio. Estos son pasos a seguir en los que no tenía experiencia o no consideraba competencia de un estadístico y que con la estancia en la Unidad comprendí la importancia que tienen.

Por lo tanto estas prácticas han sido el complemento ideal a mi formación académica, tanto por continuar con el aprendizaje metodológico estadístico como por la experiencia adquirida.

Tutores:

Dña. Carmen Cadarso Suárez
Universidade de Santiago de Compostela

Dña. Pilar Gayoso Diz
D. Francisco Gude Sampedro
Dña. María Xosé Rodríguez Álvarez

Miembros da Unidade de Epidemioloxía Clínica do Complexo Hospitalario
Universitario de Santiago de Compostela

Bibliografía

Aguilera, A. M. (2001). Tablas de contingencia bidimensionales. Hespérides La Muralla, Salamanca.

Andersen, P.K. , Borgan, Gill, R.D. y Keiding, N. (1993). Statistical Models Based on Counting Processes. N.Y. Springer-Verlag.

Bowker, A.H. (1948). A test for symmetry in contingency tables. Journal of the American Statistical Association, 43, 572-574.

Cox, D.R. (1972). Regression models and life tables (with discussion). Journal of the Royal Statistical Society: Series B, 34: 187-220.

Fisher, R.A. (1934). Statistical Methods for Research Workers. 5th Edition, Edinburgh: Oliver and Boyd.

Fleming, T.R. y Harrington, D.P. (1991). Counting Processes and Survival Analysis. N.Y.: John Wiley & Sons, Inc.

Fox, J. (2003). Effect Displays in R for Generalised Linear Models. Journal of Statistical Software, 8(15), 1-27.

Fox, J. y Hong, J (2009). Effect Displays in R for Multinomial and Proportional-Odds Logit Models: Extensions to the effects Package. Journal of Statistical Software, 32(1), 1-24.

Greenwood, M. (1926). The natural duration of cancer. Reports on Public Health and Medical Subjects, 33: 1-26, Londres: Her Majesty's Stationery Office.

Hosmer, D.W. y Lemeshow, S. (1989). Applied logistic regression. Wiley.

Kaplan, E.L. y Meier, P. (1958). Nonparametric estimation from incomplete observations. Journal of the American Statistical Association, 53: 457-481.

Klein, J.P. y Moeschberger, M.L. (1997). Survival Analysis: Techniques for Censored

Levene, H. (1960). Robust Tests for Equality of Variances, in Contributions to Probability and Statistics, ed. I. Olkin, Palo Alto, CA: Stanford Univ. Press.

Lilliefors, H.W. (1967). On the Kolmogorov-Smirnov Test for normality with mean and variance unknown. Journal of American Statistical Association, Vol 62, No.318, pp. 399-402.

Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports*, 50: 163-170.

Maxwell, A.E. (1970). Comparing the classification of subjects by two independent judges. *British journal of Psychiatry*, 116, 651-655.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12, 153-157.

Nardi, A. y Schemper, M (2003). Comparing Cox and parametric models in clinical studies. *Statistics in Medicine*, 22:3597-3610.

Peto, R. y Peto, J. (1972). Asymptotically efficient rank invariant test procedures (with discussion). *Journal of the Royal Statistical Society: Series A*, 135: 195-206.

Schoenfeld, D. (1982). Partial residuals for the proportional hazard regression model.

Shapiro, S.S. y Wilk, M.B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, Vol.52, No. 3/4, pp. 591-611.

Stuart, A. (1955). A test for homogeneity of the marginal distributions in a two-way classification. *Biometrika*, 40, 105-110.

Swets J.A. y Pickett R.M. (1982). Evaluation of diagnostic systems: methods from signal detection theory. Academic Press, Nueva York.

Therneau, T.M. y Grambsch, P.M. (2000). *Modeling Survival Data: Extending the Cox Model*. N.Y. Springer-Verlag.

Therneau, T.M., Grambsch, P.M. y Fleming, T.R. (1990). Martingale-based residuals for survival models. *Biometrika*, 77: 147-160.

Yates, F. (1934). Contingency tables involving small numbers and the χ^2 test. *Journal of the Royal Statistical Society, Suppl.1*, 217-235.

Anexo

En este anexo se muestran los scripts en R de las diferentes funciones mencionadas anteriormente. No hay que olvidar que para que se puedan ejecutar correctamente estas funciones hay que descargarse e instalar las librerías: car (para el ANOVA), nortest (para Lilliefors) y coin (para la extensión Stuart-Maxwell).

Cargar librerías

```
library(car)
library(nortest)
library(coin)
```

Script de la función “descriptiva_cuantitativa.R”

```
des.cont <- function(cont) {

  cat(paste("Número de elementos: ", length(cont)),fill=T)
  cat(paste("NAs: ", sum(is.na(cont))), fill=T)
  sum <- round(summary(cont),3)
  mean.sd <- paste(sum[4], " (", round(sd(cont, na.rm = TRUE),3), ", " ,
    round(round(sd(cont,na.rm=TRUE),3)/sqrt(length(na.omit(cont))),2)
    ,")", sep = "" )
  median.quant <- paste(sum[3], " (", sum[2], ", " , sum[5],", " " ,
    round(sum[5]-sum[2],2),")", sep = "" )
  cat(paste("Media", " (sd, se)", fill=T)
  cat(mean.sd, fill=T)
  cat("Mediana", "(1st Qu, 3rd Qu, IQR)" ,fill=T)
  cat(median.quant, fill=T)

}
```

Script de la función “descriptiva_cualitativa.R”

```
des.cat.def <- function(cat) {

  cat(paste("NAs: ", sum(is.na(cat))), fill=T)
  t <- table(cat)
  m <- matrix(ncol = 1, nrow = length(names(t)))
  rownames(m) <- names(t)
  m[,1] <- paste(t," (",round(prop.table(t)*100,2), "%)", sep = "")
  print(m, quote=F)

}
```

Script de la función “normalidad.R”

```
normalidad <- function(covar){  
  if(length(covar)>=50){  
    aux<-lillie.test(covar)  
  }else {  
    aux<-shapiro.test(covar)  
  }  
  res<-list(p.value=aux$p.value)  
  res  
}
```

Script de la función “bidimensional.R”

```
tablas.bidim<-function(cat1,cat2, datos, paired=F){  
  cat1.var<-datos[,cat1]  
  cat2.var<-datos[,cat2]  
  t<-table(cat1.var,cat2.var, dnn = c(cat1, cat2))  
  cat("*****",fill=T)  
  cat("Tabla de frecuencias absolutas",fill=T)  
  cat("*****",fill=T)  
  print(t)  
  cat("*****",fill=T)  
  cat("Tabla de frecuencias relativas",fill=T)  
  cat("*****",fill=T)  
  print(prop.table(t))  
  if(is.factor(na.omit(cat2.var))){  
    levels=levels(na.omit(cat2.var))  
  } else{  
    levels=unique(na.omit(cat2.var))  
  }  
  
  if(is.factor(na.omit(cat1.var))){  
    levels1=levels(na.omit(cat1.var))  
  } else{  
    levels1=unique(na.omit(cat1.var))  
  }  
  
  for(i in levels){  
    if(paired==F){  
      cat("*****",fill=T)  
      cat(paste("Distribución ", cat1, "/", cat2, " = ",  
        i, sep = ""),fill=T)  
      cat("*****",fill=T)  
      des.cat.def(cat1.var[cat2.var==i])  
    }  
  }  
  
  for(j in levels1){  
    if(paired==T){  
      cat("*****",fill=T)  
      cat(paste("Distribución ", cat2, "/", cat1, " = ",  
        j, sep = ""),fill=T)  
      cat("*****",fill=T)  
      des.cat.def(cat2.var[cat1.var==j])  
    }  
  }  
}
```


Script de la función “comparación_cualitativas.R”

```

test.cualitativa<- function(cat1, cat2, datos, paired=F) {
  cat1.var<-datos[,cat1]
  cat2.var<-datos[,cat2]
  cat("*****",fill=T)
  cat(paste("Descriptiva",cat1),fill=T)
  cat("*****",fill=T)
  des.cat.def(cat1.var)
  cat("*****",fill=T)
  cat(paste("Descriptiva",cat2),fill=T)
  cat("*****",fill=T)
  des.cat.def(cat2.var)
  t <- table(cat1.var,cat2.var, dnn = c(cat1, cat2))
  if(paired==F){
    if((nrow(t)==2) & (ncol(t)==2)){
      barplot(t,beside=T,ylab="Frecuencia",main="Diagrama de
        barras",legend=rownames(t),col=c(1,2),ylim=c(0,max(t)+40),
          xlab=cat2)
      tablas.bidim(cat1,cat2,datos)
      m<-c()
      esp<-matrix(,nrow(t),ncol(t))
      for(i in 1:nrow(t)){
        for(j in 1:ncol(t)){
          esp[i,j]<-((rowSums(t)[i])*(colSums(t)[j]))/sum(rowSums(t))
        }
      }
      m[i]<-sum(esp[i,]<5)
    }
    if(sum(m)==0){
      if(sum(rowSums(t))>30){
        cat("*****", fill=T)
        cat("***** TEST CHI-CUADRADO *****",fill=T)
        cat("*****", fill=T)
        print(chisq.test(cat1.var,cat2.var,correct=F))
        if(chisq.test(cat1.var,cat2.var,correct=F)$p.value>0.05){
          cat(" LAS VARIABLES SON INDEPENDIENTES",fill=T)
        } else {
          cat(" LAS VARIABLES SON DEPENDIENTES",fill=T)
        }
      }
    } else{
      cat("*****", fill=T)
      cat("***** TEST CHI-CUADRADO CORREGIDO*****",fill=T)
      cat("*****", fill=T)
      print(chisq.test(cat1.var,cat2.var))
      if(chisq.test(cat1.var,cat2.var)$p.value>0.05){
        cat(" LAS VARIABLES SON INDEPENDIENTES",fill=T)
      } else {
        cat(" LAS VARIABLES SON DEPENDIENTES",fill=T)
      }
    }
  } else{
    cat("*****", fill=T)
    cat("***** TEST DE FISHER *****",fill=T)
    cat("*****", fill=T)
    print(fisher.test(cat1.var,cat2.var))
    if(fisher.test(cat1.var,cat2.var)$p.value>0.05){
      cat(" LAS VARIABLES SON INDEPENDIENTES",fill=T)
    } else {
      cat(" LAS VARIABLES SON DEPENDIENTES",fill=T)}
  }
} else {
  barplot(t,beside=T,ylab="Frecuencia",legend=rownames(t),main="Diagrama
  de barras",col=c(1:nrow(t)),ylim=c(0,max(t)+40),xlab=cat2)
}

```

```

tablas.bidim(cat1,cat2,datos)
if(sum(rowSums(t))>30){
  cat("*****", fill=T)
  cat("***** TEST CHI-CUADRADO *****",fill=T)
  cat("*****", fill=T)
  print(chisq.test(cat1.var,cat2.var,correct=F))
  if(chisq.test(cat1.var,cat2.var,correct=F)$p.value>0.05){
    cat(" LAS VARIABLES SON INDEPENDIENTES",fill=T)
  } else {
    cat(" LAS VARIABLES SON DEPENDIENTES",fill=T)
  }
} else{
  cat("*****", fill=T)
  cat("***** TEST CHI-CUADRADO CORREGIDO *****",fill=T)
  cat("*****", fill=T)
  print(chisq.test(cat1.var,cat2.var))
  if(chisq.test(cat1.var,cat2.var)$p.value>0.05){
    cat(" LAS VARIABLES SON INDEPENDIENTES",fill=T)
  } else {
    cat(" LAS VARIABLES SON DEPENDIENTES",fill=T)
  }
}
} else {
  if((nrow(t)==2)){
    barplot(t(t),ylab="Frecuencia",main="Diagrama de
barras",legend=colnames(t),col=c(1,2),ylim=c(0,max(t)+40),
xlab=cat2)
    tablas.bidim(cat1,cat2,datos,paired=T)
    if(((t[1,2]+t[2,1])/2)<5){
      cat("*****", fill=T)
      cat("***** TEST MC-NEMAR CORREGIDO *****",fill=T)
      cat("*****", fill=T)
      print(mcnemar.test(cat1.var,cat2.var))
      if(mcnemar.test(cat1.var,cat2.var)$p.value>0.05){
        cat(" SE ACEPTA LA HIPOTESIS DE HOMOGENEIDAD
MARGINAL",fill=T)
      } else {
        cat(" SE RECHAZA LA HIPOTESIS DE HOMOGENEIDAD
MARGINAL",fill=T)
      }
    } else {
      cat("*****", fill=T)
      cat("***** TEST MC-NEMAR *****",fill=T)
      cat("*****", fill=T)
      print(mcnemar.test(cat1.var,cat2.var,correct=F))
      if(mcnemar.test(cat1.var,cat2.var,correct=F)$p.value>0.05){
        cat(" SE ACEPTA LA HIPOTESIS DE HOMOGENEIDAD
MARGINAL",fill=T)
      } else {
        cat(" SE RECHAZA LA HIPOTESIS DE HOMOGENEIDAD
MARGINAL",fill=T)
      }
    }
  }
} else {
  barplot(t(t),ylab="Frecuencia",legend=colnames(t),main="Diagrama
de barras",col=c(1:nrow(t)),ylim=c(0,max(t)+40),xlab=cat2)
  tablas.bidim(cat1,cat2,datos,paired=T)
  cat("*****", fill=T)
  cat("***** EXTENSION DE BOWKER AL TEST MC-NEMAR *****",fill=T)
  cat("*****", fill=T)
  print(mcnemar.test(cat1.var,cat2.var))
  if(mcnemar.test(cat1.var,cat2.var)$p.value>0.05){
    cat(" SE ACEPTA LA HIPOTESIS DE SIMETRIA",fill=T)
  } else {
    cat(" SE RECHAZA LA HIPOTESIS DE SIMETRIA",fill=T)
    cat("*****", fill=T)
  }
}
}

```

```

        cat("***** EXTENSION DE STUART MAXWELL AL TEST MC-NEMAR
        *****",fill=T)
        cat("*****", fill=T)
        print(mh_test(t))
    }
}
}

```

Script de la función “comparación_cuantitativas.R”

```

test.cuantitativa<- function(covar1, covar2, datos, paired=F) {
  covar1.var<-datos[,covar1]
  covar2.var<-datos[,covar2]
  cat("*****",fill=T)
  cat(paste("Descriptiva",covar1),fill=T)
  cat("*****",fill=T)
  des.cont(covar1.var)
  cat("*****",fill=T)
  cat(paste("Descriptiva",covar2),fill=T)
  cat("*****",fill=T)
  des.cont(covar2.var)
  if(paired==F){
    if(length(na.omit(covar1.var))>=50){
      cat("*****", fill=T)
      cat(paste("TEST DE LILLIEFORS para",covar1),fill=T)
      cat(paste("p-valor: ",
        round(normalidad(na.omit(covar1.var))$p.value,4)), fill=T)
      cat("*****", fill=T)
    }else {
      cat("*****", fill=T)
      cat(paste("TEST DE SHAPIRO-WILK para",covar1),fill=T)
      cat(paste("p-valor: ",
        round(normalidad(na.omit(covar1.var))$p.value,4)), fill=T)
      cat("*****", fill=T)
    }
    if(length(na.omit(covar2.var))>=50){
      cat("*****", fill=T)
      cat(paste("TEST DE LILLIEFORS para",covar2),fill=T)
      cat(paste("p-valor: ",
        round(normalidad(na.omit(covar2.var))$p.value,4)), fill=T)
      cat("*****", fill=T)
    }else {
      cat("*****", fill=T)
      cat(paste("TEST DE SHAPIRO-WILK para",covar2),fill=T)
      cat(paste("p-valor: ",
        round(normalidad(na.omit(covar2.var))$p.value,4)),fill=T)
      cat("*****", fill=T)
    }
  }
  if((normalidad(na.omit(covar1.var))$p.value>0.05) &
    (normalidad(na.omit(covar2.var))$p.value>0.05)){
    cat("*****", fill=T)
    cat("***** TEST PARAMÉTRICOS EN POBLACIONES NORMALES
    *****",fill=T)
    cat("*****", fill=T)
    cat("***** Comparación de varianzas *****",fill=T)
    print(var.test(covar1.var,covar2.var))
    cat(paste("p-valor: ", var.test(covar1.var,covar2.var)$p.value),
      fill=T)
    if(var.test(covar1.var,covar2.var)$p.value<0.05){
      cat(" NO TIENEN LA MISMA VARIANZA",fill=T)
      cat("***** Comparación de medias *****",fill=T)
      print(t.test(covar1.var,covar2.var))
      if(t.test(covar1.var,covar2.var)$p.value<0.05){

```

```

        cat(" NO TIENEN LA MISMA MEDIA",fill=T)
    } else {
        cat(" TIENEN LA MISMA MEDIA",fill=T)
    }
} else {
    cat(" TIENEN LA MISMA VARIANZA",fill=T)
    cat("***** Comparación de medias *****",fill=T)
    print(t.test(covar1.var, covar2.var, var.equal=T))
    if(t.test(covar1.var, covar2.var, var.equal=T)$p.value<0.05){
        cat(" NO TIENEN LA MISMA MEDIA",fill=T)
    } else {
        cat(" TIENEN LA MISMA MEDIA",fill=T)
    }
}
} else {
    cat("*****", fill=T)
    cat("***** TEST NO PARAMÉTRICOS PARA DOS MUESTRAS *****",fill=T)
    cat("*****", fill=T)
    cat("***** Test de Wilcoxon *****",fill=T)
    print(wilcox.test(covar1.var, covar2.var))
    if(wilcox.test(covar1.var, covar2.var)$p.value<0.05){
        cat(" NO TIENEN LA MISMA DISTRIBUCIÓN",fill=T)
    } else {
        cat(" TIENEN LA MISMA DISTRIBUCIÓN",fill=T)
    }
}
} else {
    if(length(na.omit(covar1.var))>=50){
        cat("*****", fill=T)
        cat(paste("TEST DE LILLIEFORS para", covar1), fill=T)
        cat(paste("p-valor: ",
            round(normalidad(na.omit(covar1.var))$p.value, 4)), fill=T)
        cat("*****", fill=T)
    }else {
        cat("*****", fill=T)
        cat(paste("TEST DE SHAPIRO-WILK para", covar1), fill=T)
        cat(paste("p-valor: ",
            round(normalidad(na.omit(covar1.var))$p.value, 4)), fill=T)

        cat("*****", fill=T)
    }
    if(length(na.omit(covar2.var))>=50){
        cat("*****", fill=T)
        cat(paste("TEST DE LILLIEFORS para", covar2), fill=T)
        cat(paste("p-valor: ",
            round(normalidad(na.omit(covar2.var))$p.value, 4)), fill=T)
        cat("*****", fill=T)
    }else {
        cat("*****", fill=T)
        cat(paste("TEST DE SHAPIRO-WILK para", covar2), fill=T)
        cat(paste("p-valor: ",
            round(normalidad(na.omit(covar2.var))$p.value, 4)), fill=T)
        cat("*****", fill=T)
    }
    if((normalidad(na.omit(covar1.var))$p.value>0.05) &
        (normalidad(na.omit(covar2.var))$p.value>0.05)){
        cat("*****", fill=T)
        cat("***** TEST PARAMÉTRICOS EN POBLACIONES NORMALES
            *****",fill=T)
        cat("*****", fill=T)
        cat("***** Comparación de medias para muestras relacionadas
            *****",fill=T)
        print(t.test(covar1.var, covar2.var, paired=TRUE))
        if(t.test(covar1.var, covar2.var, paired=TRUE)$p.value<0.05){
            cat(" NO TIENEN LA MISMA MEDIA",fill=T)
        } else {
            cat(" TIENEN LA MISMA MEDIA",fill=T)
        }
    }
}
}

```

```

} else {
  cat("*****", fill=T)
  cat("***** TEST NO PARAMÉTRICOS *****", fill=T)
  cat("*****", fill=T)
  cat("***** Test de Wilcoxon para muestras apareadas
      *****", fill=T)
  print(wilcox.test(covar1.var, covar2.var, paired=T))
  if(wilcox.test(covar1.var, covar2.var, paired=T)$p.value<0.05) {
    cat(" NO TIENEN LA MISMA DISTRIBUCIÓN", fill=T)
  } else {
    cat(" TIENEN LA MISMA DISTRIBUCIÓN", fill=T)
  }
}
}
l=list(covar1=covar1.var, covar2=covar2.var)
names(l)=c(covar1, covar2)
windows()
boxplot(l, main="Diagrama de Cajas", col="red")
windows()
plot(density(na.omit(covar1.var)), main="Comparacion de
      Densidades", ylim=c(0, max(max(density(na.omit(covar1.var))$y), max(density
      (na.omit(covar2.var))$y))+0.01))
lines(density(na.omit(covar2.var)), col=2)
legend("topleft", legend=names(l), col=c(1,2), box.lwd=1, lty=c(1,1))
}

```

Script de la función “comparación_mixta.R”

```

test.mixto <- function(cont, cat, datos, paired=F) {
  cont.var<-datos[,cont]
  cat.var<-datos[,cat]
  cat("*****", fill=T)
  cat(paste("Descriptiva", cont), fill=T)
  cat("*****", fill=T)
  des.cont(cont.var)
  cat("*****", fill=T)
  cat(paste("Descriptiva", cat), fill=T)
  cat("*****", fill=T)
  des.cat.def(cat.var)

  if(is.factor(cat.var)) {
    levels=levels(cat.var)
  } else {
    levels=unique(cat.var)
  }
  windows()
  boxplot(cont.var~cat.var, main="Diagrama de Cajas",
          xlab=cat, ylab=cont, col="red")
  if(paired==F) {
    if( nrow(table(cat.var))==2) {
      windows()
      plot(density(na.omit(cont.var[cat.var==levels[1]])), main=
            "Estimación de densidades \n según variable
            categórica", ylim=c(0, max(max((density(na.omit(cont.var[cat.
            var==levels[1]))$y), max((density(na.omit(cont.var[cat.var
            ==levels[2]))$y))+0.01))
      lines(density(na.omit(cont.var[cat.var==levels[2]])), col=2)
      legend("topleft", legend=levels, col=c(1,2), box.lwd=1, lty=c(1,1))
      for (i in levels) {
        cat("*****", fill=T)
        cat(paste("categórica-nivel:", i), fill=T)
        cat("*****", fill=T)
        des.cont(cont.var[cat.var==i])
        if(length(na.omit(cont.var[cat.var==i]))>=50) {

```



```

legend("topleft", legend=levels, col=1:j, box.lwd=1, lty=c(1,1))
n<-c()
for (i in levels){
  cat("*****", fill=T)
  cat(paste("categórica-nivel:", i), fill=T)
  cat("*****", fill=T)
  des.cont(cont.var[cat.var==i])
  if(length(na.omit(cont.var[cat.var==i]))>=50){
    cat("*****", fill=T)
    cat(paste("TEST DE LILLIEFORS para nivel", i), fill=T)
    cat(paste("p-valor: ", round(normalidad(na.omit(cont.var[cat.var==i]))$p.value,4), fill=T)
    cat("*****", fill=T)
  }else {
    cat("*****", fill=T)
    cat(paste("TEST DE SHAPIRO-WILK para nivel", i), fill=T)
    cat(paste("p-valor: ", round(normalidad(na.omit(cont.var[cat.var==i]))$p.value,4), fill=T)
    cat("*****", fill=T)
  }
  n[i]<-
  normalidad(na.omit(cont.var[cat.var==i]))$p.value<0.05
}
if(sum(n)==0){
  cat("***** Comparación de varianzas entre grupos", fill=T)
  print(leveneTest(cont.var~as.factor(cat.var)))
  if(leveneTest(cont.var~as.factor(cat.var))[3][1,1]>0.05){
    cat("***** Las hipótesis del modelo ANOVA son válidas", fill=T)
    cat("***** ANOVA", fill=T)
    if(length(na.action(na.omit(cont.var)))==0){
      print(summary(aov(cont.var~as.factor(cat.var))))
      if(summary(aov(cont.var~as.factor(cat.var)))[1][[1]][1,5]>0.05){
        cat("***** TIENEN LA MISMA MEDIA", fill=T)
      }else {
        cat("***** NO TIENEN LA MISMA MEDIA", fill=T)
      }
      cat("***** Método de la diferencia significativa de Tukey", fill=T)
      HSD<-
      TukeyHSD(aov(cont.var~as.factor(cat.var)), conf.level=0.95)
      plot(HSD)
    }
  } else {
    Motivos<-
    cat.var[-na.action(na.omit(cont.var))]
    print(summary(aov(na.omit(cont.var)~as.factor(Motivos))))
    if(summary(aov(na.omit(cont.var)~as.factor(Motivos)))[1][[1]][1,5]>0.05){
      cat("***** TIENEN LA MISMA MEDIA", fill=T)
    }else {
      cat("***** NO TIENEN LA MISMA MEDIA", fill=T)
      cat("***** Método de la diferencia significativa de Tukey", fill=T)
    }
  }
}

```

```

HSD<-
  TukeyHSD(aov (na.omit (cont.var) ~a
s.factor (Motivos)), conf.level=0.
95)
plot (HSD)
}
}
} else{
  cat("***** Las hipótesis del modelo ANOVA no son
válidas *****", fill=T)
  cat("*****", fill=T)
  cat("***** Test de Kruskal-Wallis para k muestras
independientes *****", fill=T)
  cat("*****", fill=T)
  print(kruskal.test (cont.var~cat.var))
  if(kruskal.test (cont.var~cat.var) $p.value<0.05) {
    cat(" NO TIENEN LA MISMA
DISTRIBUCIÓN", fill=T)
  } else{
    cat(" TIENEN LA MISMA DISTRIBUCIÓN", fill=T)
  }
}
} else{
  cat("*****", fill=T)
  cat("***** Test de Kruskal-Wallis para k muestras
independientes *****", fill=T)
  cat("*****", fill=T)
  print(kruskal.test (cont.var~cat.var))
  if(kruskal.test (cont.var~cat.var) $p.value<0.05) {
    cat(" NO TIENEN LA MISMA DISTRIBUCIÓN", fill=T)
  } else {
    cat(" TIENEN LA MISMA DISTRIBUCIÓN", fill=T)
  }
}
}
}
if (paired==T) {
  if ( nrow (table (covar)) ==2) {
    windows ()
    plot (density (na.omit (cont.var [cat.var ==levels [1]])), main="
Estimación de densidades \n según variable
categórica", ylim=c (0, max (max (density (na.omit (cont.var [cat.
var ==levels [1]])) $y), max (density (na.omit (cont.var [vat.var
==levels [2]])) $y)) +0.01))
    lines (density (na.omit (cont.var [cat.var ==levels [2]])), col=2)
    legend ("topleft", legend=levels, col=c (1,2), box.lwd=1, lty=c (1,1))
    for (i in levels) {
      cat (paste ("categórica-nivel:", i), fill=T)
      des.cont (cont.var [cat.var ==i])
      if (length (na.omit (cont.var [cat.var ==i])) >=50) {
        cat ("*****", fill=T)
        cat (paste ("TEST DE LILLIEFORS para nivel
=", i), fill=T)
        cat (paste ("p-valor: ",
round (normalidad (na.omit (cont.var [cat.var ==i]
)) $p.value, 4), fill=T)
        cat ("*****", fill=T)
      } else {
        cat ("*****", fill=T)
        cat (paste ("TEST DE SHAPIRO-WILK para nivel
=", i), fill=T)
        cat (paste ("p-valor: ",
round (normalidad (na.omit (cont.var [cat.var ==i]
)) $p.value, 4), fill=T)
      }
    }
  }
}
if ((normalidad (na.omit (cont.var [cat.var ==levels [1]])) $p.value >0.05) &
(normalidad (na.omit (cont.var [cat.var ==levels [2]])) $p.value >0.05)) {
  cat ("*****", fill=T)

```



```

cat("***** TEST PARAMÉTRICOS EN POBLACIONES NORMALES
*****",fill=T)
cat("*****", fill=T)
cat("***** Comparación de medias *****",fill=T)
print(t.test(cont.var~cat.var,paired=T))
if(t.test(cont.var~cat.var,paired=T)$p.value<0.05){
  cat(" NO TIENEN LA MISMA MEDIA",fill=T)
} else {
  cat(" TIENEN LA MISMA MEDIA",fill=T)
}
} else {
cat("*****", fill=T)
cat("***** TEST NO PARAMÉTRICOS PARA DOS MUESTRAS *****",fill=T)
cat("*****", fill=T)
cat("***** Test de Wilcoxon para dos muestras apareadas
*****",fill=T)
print(wilcox.test(cont.var~cat.var,paired=T))
if(wilcox.test(cont.var~cat.var,paired=T)$p.value<0.05){
  cat(" NO TIENEN LA MISMA DISTRIBUCIÓN",fill=T)
} else {
  cat(" TIENEN LA MISMA DISTRIBUCIÓN",fill=T)
}
}
} else {
windows()
plot(density(na.omit(cont.var[cat.var==levels[1]]),main="Estimación de
densidades \n según variable categórica"
,ylim=c(0,max(density(na.omit(marker[covar==1]))$y)+0.03))
j=1
for (i in levels[-1]){
  j=j+1
  lines(density(na.omit(cont.var[cat.var==i])),col=j)
}
legend("topleft",legend=levels,col=j,box.lwd=1,lty=c(1,1))
for (i in levels){
  cat(paste("categórica-nivel:",i),fill=T)
  des.cont(cont.var[cat.var==i])
}
if(length(na.action(na.omit(cont.var)))==0){
  cat("*****", fill=T)
  cat("***** Test de Friedman para k muestras relacionadas
*****",fill=T)
  cat("*****", fill=T)
  print(friedman.test(cont.var~cat.var))
  if(friedman.test(cont.var~cat.var)$p.value<0.05){
    cat(" NO TIENEN LA MISMA DISTRIBUCIÓN",fill=T)
  } else {
    cat(" TIENEN LA MISMA DISTRIBUCIÓN",fill=T)
  }
}
} else{
Motivos<-as.factor(cat.var)[-na.action(na.omit(cont.var))]
cat("*****", fill=T)
cat("***** Test de Friedman para k muestras relacionadas
*****",fill=T)
cat("*****", fill=T)
print(friedman.test(na.omit(cont.var)~Motivos))
if(friedman.test(na.omit(cont.var)~Motivos)$p.value<0.05){
  cat(" NO TIENEN LA MISMA DISTRIBUCIÓN",fill=T)
} else{
  cat(" TIENEN LA MISMA DISTRIBUCIÓN",fill=T)
}
}
}
}
}
}

```