UNIVERSIDAD DE SANTIAGO DE CHILE
FACULTAD DE INGENIERÍA
DEPARTAMENTO DE INGENIERÍA INFORMÁTICA

ANÁLISIS DE DATOS

LABORATORIO 1

BREAST CÁNCER WISCONSIN

YANIRA SÁEZ
DIEGO SALINAS

Profesor:
Sr. Felipe Bello
Ayudantes:
Sr. Brayan Guzmán
Srta. Fernanda Lobos

Santiago - Chile
6 de septiembre de 2016

TABLA DE CONTENIDOS

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

CAPÍTULO 1. INTRODUCCIÓN

Los primeros indicios del cáncer de mama datan cerca del año 1600 a.C en Egipto, pero no fue hasta el siglo XVII que se lograron verdaderos avances como el comienzo de las tan conocidas mastectomias, un procedimiento muy común en los años 1970 para erradicarlo. A pesar de todos estos intentos, el cáncer de mama continúa siendo el cáncer más común y la segunda causa de muerte por canceres en mujeres, en los últimos 20 años, 28 de cada 100 mujeres mueren a causa de este tipo de cáncer. Por esta razón, tanto el mundo científico en conjunto con la sociedad dan tanto énfasis en campañas preventivas para su pronta detección, ya que detectando a tiempo esta enfermedad y utilizando el tratamiento adecuado según las características del tumor del paciente, ya sea benigno o maligno, se puede prolongar los años y la calidad de vida del paciente.[**1**]

A partir de la base de datos denominada "Breast Cancer Wisconsin" donada por el académico Olvi Mangasarian de la Universidad de Wisconsin, además de los papers relacionados con dicha base de datos, se tiene como objetivo principal en esta laboratorio estudiar y analizar los datos correspondientes a la base de datos mencionada anteriormente.Por lo tanto, es necesario cumplir ciertos objetivos específicos descritos a continuación:

- Estudiar los conceptos relacionados con el área de la salud, para lograr una mayor comprensión del problema.
- Describir la información relevante de la base de datos.
- Definir y comprender el significado de clases, atributos y sus valores.
- Utilizar el software R para los cálculos estadísticos.
- Realizar un análisis completo al problema planteado según los conocimientos estadísticos e inferenciales adquiridos en clases.

La estructura del informe se divide en una pequeña Introducción, seguida del capítulo de Descripción del Problema el cual presenta una descripción de la base de datos, de las clases y variables utilizadas, en el tercer capítulo se realiza un Análisis Estadísticos e Inferencial según el dato utilizado, luego el capítulo cuatro de las Conclusiones pertinentes al desarrollo y resultado obtenidos, finalmente la Bibliografía utilizada para el correcto desarrollo del laboratorio.

CAPÍTULO 2. DESCRIPCIÓN DEL PROBLEMA

El Cáncer es el crecimiento descontrolado de células anormales en el cuerpo. Las células cancerosas también se denominan células malignas.El cáncer se origina de células en el cuerpo. Las células normales se multiplican cuando el cuerpo las necesita y mueren cuando se dañan o cuando el cuerpo ya no las necesita.

Este parece ocurrir cuando el material genético de una célula cambia. Eso provoca que las células crezcan fuera de control. Las células se dividen demasiado rápido y no mueren de la manera normal.

Existen muchos tipos diferentes de cáncer. Puede aparecer en casi cualquier órgano o tejido, como el pulmón, el colon, los senos, la piel, los huesos o el tejido nervioso. Existen múltiples factores de riesgo para el cáncer como:

- Benceno y otros químicos.
- Beber demasiado alcohol.
- Toxinas ambientales, como ciertos hongos venenosos y un tipo de tóxico que puede formarse en las plantas de cacahuete (aflatoxinas).
- Problemas genéticos.
- Obesidad.
- Exposición a la radiación.
- Demasiada exposición al sol.
- Virus.

Aunque la causa de muchos tipos de cáncer aún sigue siendo desconocida.
En los hombres estadounidenses, más allá del cáncer de piel, los tres cánceres más comunes son:

- Cáncer de próstata.
- Cáncer pulmonar.
- Cáncer colorrectal.

En las mujeres estadounidenses, más allá del cáncer de piel, los tres cánceres más comunes son:

- Cáncer de mama.
- Cáncer pulmonar.
- Cáncer colorrectal.[**2**]

El cáncer de mama es uno de los canceres más comunes en las mujeres, y la segunda causa de muerte por cáncer en mujeres, es una realidad lamentable a la cual se enfrenta nuestra sociedad, pero es por dichos datos que personas como el Doctor William H. Wolberg han puesto sus conocimientos y tiempos en investigar métodos para la pronta detección de dicha enfermedad, ya que con una detección temprana y un tratamiento adecuado a las características que causan el cáncer del paciente se puede mejorar sus pronósticos de calidad de vida.

2.1. Descripción de la Base de Datos

La base de datos Wisconsin Diagnostic Breast Cancer (WDBC) fue recolectada y creada por los académicos de la Universidad de Wisconsin el Dr.William H. Wolberg, W. Nick Street y Olvi L. Mangasarian, y donada a Machine Learning Repository el 15 de julio de 1992. La base de datos esta compuesta por 699 muestras de pacientes que se les detectaron tumores, 458 tumores benignos y 241 tumores malignos.

Las características se calculan a partir de una imagen digitalizada de un aspirado con aguja fina (FNA) de un bulto en la mama. Se describen las características de los núcleos de las células presentes en la imagen. La separación del plano descrito anteriormente se obtuvo utilizando el Método árbol-multisuperficies (HSH-T), un método de clasificación que utiliza la programación lineal para construir un árbol de decisión. Las características relevantes fueron seleccionadas mediante una búsqueda exhaustiva en el espacio de 1-4 y 1-3 características que separan los planos. [**3**]

Características de la base de datos:

- **Características del conjunto de datos:** Multivariables.
- **Número de instancias:** 699.
- **Número de atributos:** 10 más el atributo Clase.
- **Características de los atributos:** Natural ie. $atributo_i \in \mathbb{N}$.
- **Objetivo de la muestra:** Clasificación.
- **Valores perdidos:** 16.

La base de datos se compone de los siguientes archivos:

- **breast-cancer-wisconsin.csv:** Es un archivo con el formato CSV, que es un tipo de documento en formato abierto sencillo para representar datos en forma de tabla, en las que las columnas se separan por comas y las filas por saltos de línea.[**4**] Este archivo contiene las 699 muestras de los pacientes y 10 de las 11 variables, exceptuando el número de código de la muestra.

- **breast-cancer-wisconsin.data:** Es un archivo en formato DAT, o también denominado ficheros de datos, este tipo de archivos contienen datos genéricos que pueden ser utilizados o indexados por otros programas. Este archivo contiene todos los valores para 11 variables existentes en la base de datos, las cuales se presentan en el siguiente orden: Número de código de la muestra,Grupo de grosor, Tamaño uniforme de la célula, Forma uniforme de la célula, Adhesión marginal, Tamaño único de células epiteliales, Núcleo justo, Cromatina, Nucléolo normal, Mitosis, Clase.

- **breast-cancer-wisconsin.names:** Es un archivo con extensión .names, este tipo de archivos describe el nombre y tipo de los campos o atributos del archivo de dato, En este archivo se encuentra información relevante como: papers en los cuales se ha utilizado la base de datos, sus creadores y donadores, los grupos de estudio, el número de instancias y atributos y la distribución de la variable Clase.

2.2. Descripción de las clases y variables

La base de datos esta compuesta por 11 variables seleccionadas que ayudan a distinguir entre tumores benignos y malignos, se calificaron de 1 a 10 (el valor 1 correspondiendo a un estado normal y 10 a un estado más anormal) en el momento de recolección de muestras de las 9 variables que aportan información relativa a la naturaleza de la observación, por otra parte la variable Clase que informa sobre la clasificación de estas, quien puede adoptar el valor 2 o 4 y la variable Número de código de la muestra que es un ID [**5**].A continuación se definen y describen las clases y variables utilizadas en la base de datos para su posterior análisis.

1. **Número de código de la muestra:** Identificador de cada paciente. Es una variable de tipo cuantitativa discreta.

2. **Grupo de grosor (clump):** Células benignas tienden a agrupar en monocapas, mientras las células cancerosas se agrupan a menudo en multicapas.

3. **Tamaño uniforme de la célula (size):** Se evalúa la consistencia en el tamaño de las células de la muestra. En las células cancerosas tienden a variar en tamaño. Es por esto que este parámetro es importante para determinar si una célula es cancerosa o no.

4. **Forma uniforme de la célula (shape):** Se estima la igualdad en la forma de las células e identifica varianzas marginales. En las células cancerosas tienden a variar su forma.

5. **Adhesión marginal (adhesion):** Las células normales tienden a permanecer juntas. Las células cancerosas tienden a perder esta habilidad. Entonces si pierden la adhesión es un signo de tumor maligno.

6. **Tamaño único de células epiteliales (epithelial):** Como se mencionó la uniformidad anteriormente. Las células epiteliales que son significativamente más grandes pueden ser células malignas.

7. **Núcleo justo (nuclei):** Este término es utilizado para núcleos que no están rodeados por el citoplasma (el resto de la célula). Esto es típico en tumores benignos.

8. **Cromatina (chromatin):** Describe una "textura" uniforme de los núcleos en células benignas. En las células cancerosas la cromatina tiende a ser gruesa.

9. **Nucléolo normal (nucleoli):** Los nucléolos son pequeñas estructuras vistas en el núcleo. En las células normales los nucléolos normalmente son muy pequeños si es visible en absoluto. En las células cancerosas los nucléolos son más prominentes y a veces hay más de ellos.

10. **Mitosis (mitoses):** Describe el nivel de actividad mitótica (reproducción celular).

11. **Clase (class):** Clasifica la clase de los tumores en dos tipos, asignando el número 2 para benigno y 4 para maligno. Es un tipo de variables cualitativa ordinal, debido a que admite un criterio de orden ya que va desde 2 para benigno y luego 4 para maligno.[**6**]

CAPÍTULO 3. ANÁLISIS ESTADÍSTICO E INFERENCIAL

En este capítulo se muestran los resultados obtenidos mediante los cálculos realizados con el software R y sus respectivos análisis para una mayor comprensión del problema planteado en el capítulo anterior.

De las 699 muestras que proporciona la base de datos, hay 16 instancias perdidas representadas con un signo *?* en el atributo *nuclei*, por lo que se decidió eliminar los 16 objetos que los contienen para no entorpecer ciertos cálculos lo cual tiene como consecuencia una perdida del 2,28 % de información para el estudio, considerada aceptable.

Por lo tanto, quedaron 683 muestras para realizar el estudio, correspondiendo 444 a tumores benignos y 239 a tumores malignos, siendo calificadas en un rango de 1 a 10 cada atributo como se mencionó en el capítulo anterior.

En la tabla 3.1 se muestran las medidas de tendencia central y variación, como se puede observar la moda en la mayoría de los atributos es 1 al igual que la mediana , por lo tanto los datos concuerdan con que la mayoría de las muestras son tumores benignos correspondientes al 65 %, también considerable con respecto a la media, en donde casi todos de los atributos tienen una clasificación con tendencia promedio a la normalidad, con excepción de clump, cuya media está en casi el centro de la distancia entre normal y anormal.

Los atributos que poseen una relativa baja varianza son clump, size, shape, epithelial y chromatin; y los que poseen una relativa alta varianza son adhesion, nuclei,nucleoli y mitoses, lo que nos indica cuan homogéneo son los datos y nos dá un indicio de cómo es la curtosis de la distribución que las representa, a menor varianza, mayor curtosis, mayor homogeneidad y viceversa.

Tabla 3.1: Medidas de tendencia central y variación

Variable \ Medida	Moda	Mediana	Media	Varianza	Coef variación	Rango
clump	1	4	4,442167	7,956694	0,6349967	$[1-10]$
size	1	1	3,150805	9,395113	0,9728132	$[1-10]$
shape	1	1	3,215227	8,931615	0,9295085	$[1-10]$
adhesion	1	1	2,830161	8,205717	1,0121552	$[1-10]$
epithelial	2	2	3,234261	4,942109	0,6873551	$[1-10]$
nuclei	1	1	3,544656	13,277695	1,0279861	$[1-10]$
chromatin	3	3	3,445095	6,001013	0,7110679	$[1-10]$
nucleoli	1	1	2,869693	9,318772	1,0637608	$[1-10]$
mitoses	1	1	1,603221	3,002160	1,0807456	$[1-10]$

El coeficiente de variación nos indica la proporción en la variación de los datos con respecto a la media, cuando tiende a 0 implica que la variación de las instancias con respecto a la media del atributo es poca, en cambio cuando tiende a 1 implica lo opuesto. Es importante aclarar que cuando la media tiende a 0, el coeficiente no entrega mucha información debido a que el aumento de este no implica la dispersión de los datos. Por otra parte, como se observa en la Tabla 3.1, los atributos adhesion, nuclei, nucleoli y mitoses presentan un coeficiente de variación superior a 1, casos no tan comunes, quienes pueden ser explicados por distribuciones que lo permiten.

La Tabla 3.2 muestra las frecuencias absolutas de los grados de clasificación que van desde 1 a 10 representando el estado normal(1) o más anormal(10) de la característica del tumor de sus correspondientes 9 atributos. Se puede observar que en la mayoría de las variables los tres primeros valores de grado poseen mayor frecuencia, lo cual concuerda con las 444 muestras correspondientes a tumores benignos debido a su clasificación como Normal.Por otro lado las tres variables que podrían tener mayor influencia en los 239 muestras de tumores malignos debido a obtener las mayores frecuencias en el estado anormal serían núcleo, grosor y tamaño.

Tabla 3.2: Frecuencias Absolutas de las Variables

Variable \ Grado	1	2	3	4	5	6	7	8	9	10
clump	139	50	104	79	128	33	23	44	14	69
size	373	45	52	38	30	25	19	28	6	67
shape	346	58	53	43	32	29	30	27	7	58
adhesion	393	58	58	33	23	21	13	25	4	55
epithelial	44	376	71	48	39	40	11	21	2	31
nuclei	402	30	28	19	30	4	8	21	9	132
chromatin	150	160	161	39	34	9	71	28	11	20
nucleoli	432	36	42	18	19	22	16	23	15	60
mitoses	563	35	33	12	6	3	9	8	0	14

Como se muestra en la Figura 3-1 quien dispone de los 9 atributos, en donde cada atributo es graficado según la frecuencia absoluta de el grado de normalidad y a su vez es separado en porciones, vale decir, nos muestra en que proporción el grado está presente en la clase del tumor (benigno 2, maligno 4).
En todos los gráficos se puede apreciar que a medida que el grado pasa de normal a anormal (de 1 a 10) la proporción dominante del atributo presente en la clase pasa de benigno a maligno, en general, las instancias de los primeros 3 grados tienen una alta probabilidad de pertenecer a objetos no cancerígenos. Caso particular con el atributo mitoses quien posee una mayor probabilidad sólo para el primer grado de normalidad.

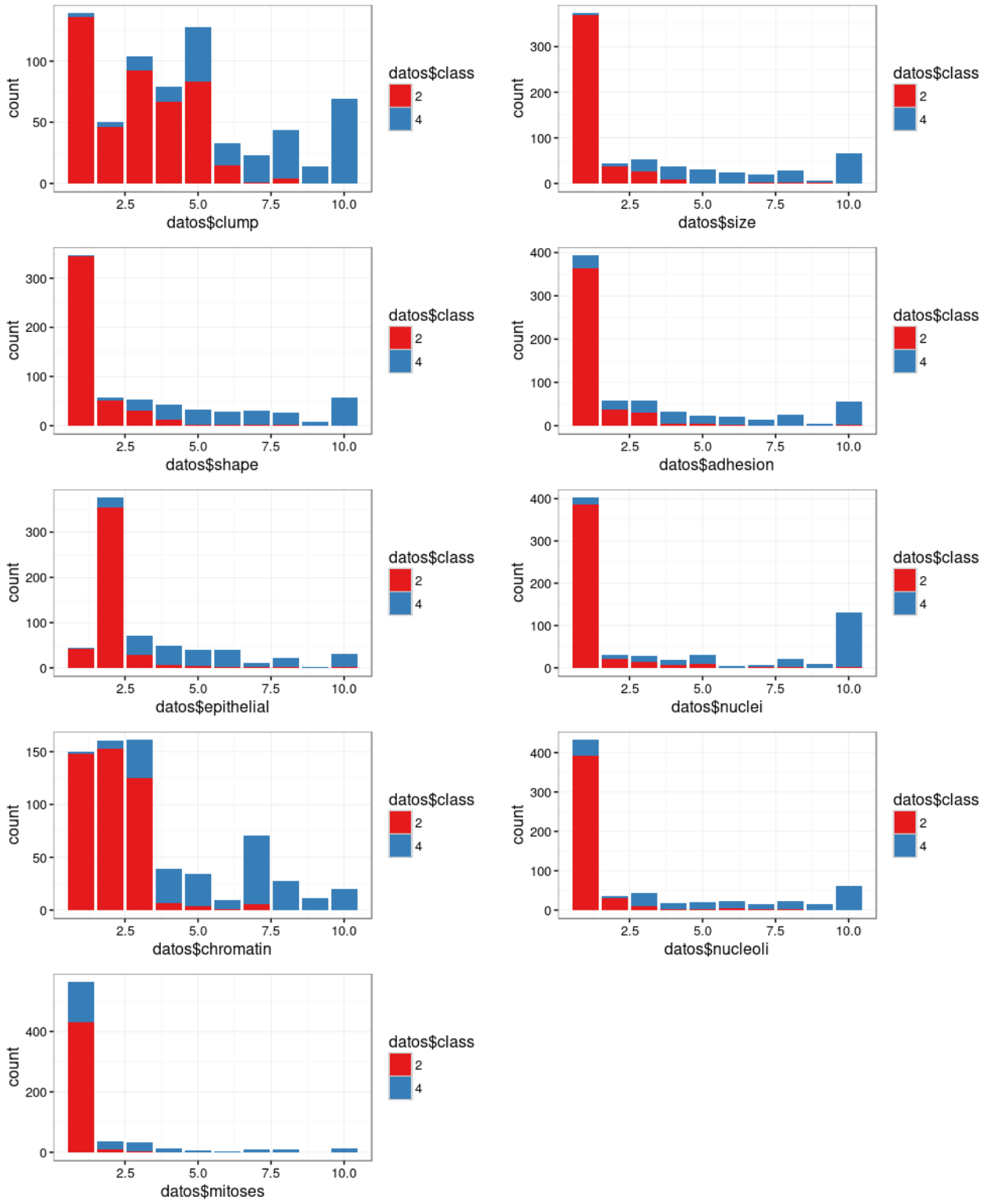

Figura 3-1: Proporción de Grado con respecto a la Clase de los Atributos

Como no se conoce la distribución de los atributos, haremos uso de estadística no paramétrica para ahondar sobre su comportamiento. En una primera instancia con la función *descdist()* para cada uno de los atributos, con el parámetro *discrete =TRUE* debido a que las escalas corresponden a valores enteros, es posible obtener un gráfico de Cullen and Frey, el que muestra a qué distribución se asemeja más el conjunto de datos entregado como parámetro, como se puede apreciar en la Figura 3-2 como ejemplo para el atributo clump, en donde el punto azul indica la cercanía de la observación del atributo con las distintas distribuciones discretas de prueba, en este caso Normal, Poisson y Binomial Negativa. Es importante aclarar que este gráfico es generado para todos los atributos (donde se observa que tienen como posibilidad las mismas distribuciones discretas que el ejemplo), quienes en mayor o menor grado, muestran similitudes para algunas de las 3 distribuciones (normal,poisson, binomial negativa), por lo que es el punto de partida para la siguiente prueba.(Para mayor detalle correr el script adjuntado en el Anexo)

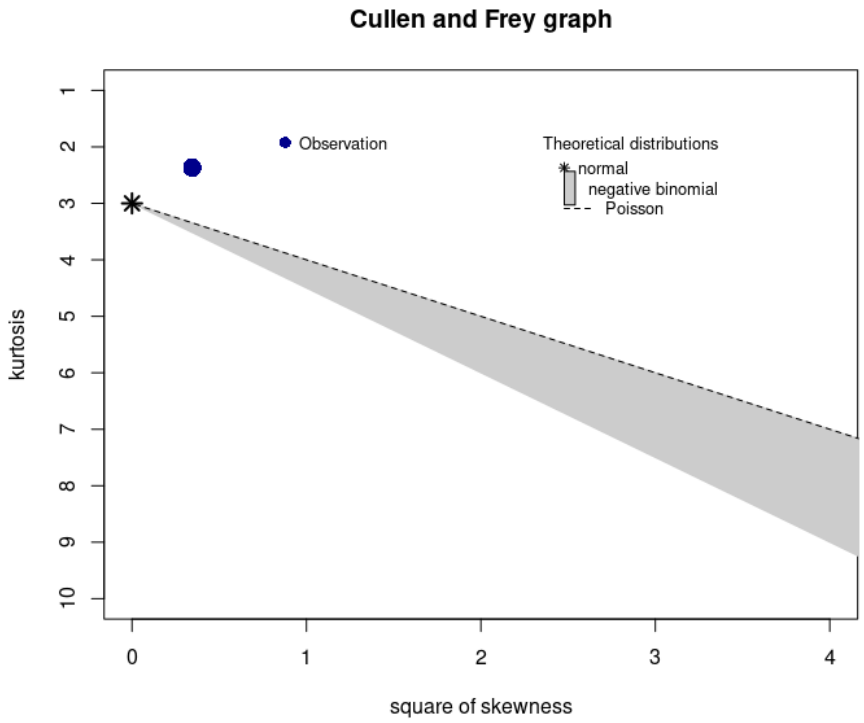

Figura 3-2: Cullen and Frey Graph para Clump

con la función *fitdist()* quien busca el ajuste del atributo con la distribución de prueba (en este caso, para las 3 obtenidas anteriormente, también para cada atributo) por el método de 'maximum likelihood', luego de esto, se procede a contrastar las distribuciones con el conjunto de prueba como se muestra en la Figura 3-3 como ejemplo para el atributo clump, donde se puede observar que la distribución de prueba cuya diferencia con la distribución del atributo es mas parecida es la de Binomial Negativa, vale decir, es la que a simple vista, posee mejor ajuste.

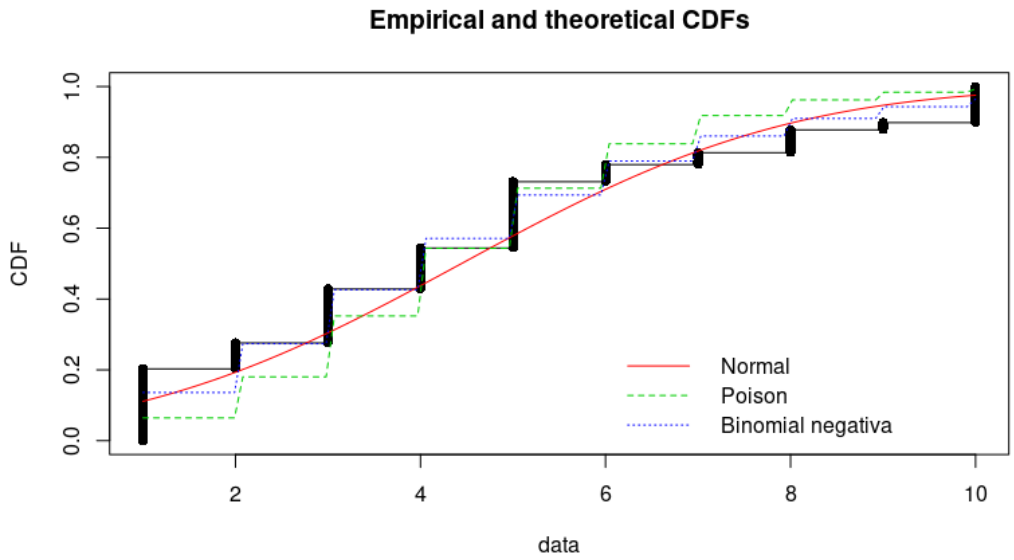

Figura 3-3: Contraste distribuciones para atrituto Clump

Para Obtener una aproximación mas cuantificable, utilizaremos una prueba de bondad de ajuste que nos permita dilucidar la distribución que mejor represente cada atributo

A continuación, con la ayuda de la función *gofstat()* se procede a realizar la prueba de hipótesis para funciones discretas χ^2 (Chi cuadrado), quien nos entrega el estadístico correspondiente a las diferencias de ajuste entre ambas distribuciones, teórica y empírica, por lo que a menor estadístico mejor es el ajuste. Por otra parte, esta función nos otorga 2 criterios de información importantes:.*El AIC que fue propuesto por Akaike (1974) como un estimador insesgado asintótico de la información de Kullback-Leibler esperada, entre un modelo candidato ajustado y el verdadero modelo;El BIC que fue derivado por Schwarz en 1978 como una aproximación a una transformación de la probabilidad posterior de un modelo candidato."* [**7**] Por esto a menor BIC implica un número menor de variables explicativas, mejor ajuste y por otro lado a menor AIC implica una menor perdida de información, así entonces se puede observar la información resumida en la Tabla 3.3, en la que se Acepta el mejor (menor) estadístico o criterio para los distintos atributos según sea el caso, ya que se busca minimizar la diferencia entre el atributo y la distribución a elegir, maximizando el ajuste y minimizando la pérdida de información.

Tabla 3.3: Prueba Gofstat: Chi cuadrado, AIC y BIC

variable	Distribucion / Test	Normal	Poisson	Bin Neg	Aceptada
	Chi-squared statistic	172,4774	456,8575	168,3217	Bin Neg
Clump	Aikake's Information Criterion	3357,821	3393,933	3251,288	Bin Neg
	Bayesian Information Criterion	3366,874	3398,459	3260,341	Bin Neg
	Chi-squared statistic	711,8172	6450,233	597,779	Bin Neg
Size	Aikake's Information Criterion	3471,319	3568,377	3065,370	Bin Neg
	Bayesian Information Criterion	3480,372	3572,904	3074,423	Bin Neg
	Chi-squared statistic	575,1984	4370,248	447,0667	Bin Neg
Shape	Aikake's Information Criterion	3436,764	3511,728	3072,634	Bin Neg
	Bayesian Information Criterion	3445,817	3516,255	3081,687	Bin Neg
	Chi-squared statistic	478,4366	2581,153	294,4227	Bin Neg
Adhesion	Aikake's Information Criterion	3378,869	3356,408	2929,835	Bin Neg
	Bayesian Information Criterion	3387,922	3360,935	2938,888	Bin Neg
	Chi-squared statistic	478,4366	2581,153	294,4227	Bin Neg
Ephitelial	Aikake's Information Criterion	3378,869	3356,408	2929,835	Bin Neg
	Bayesian Information Criterion	3387,922	3360,935	2938,888	Bin Neg
	Chi-squared statistic	743,8386	4213,28	765,4624	Normal
Nuclei	Aikake's Information Criterion	3707,566	4119,105	3250,203	Bin Neg
	Bayesian Information Criterion	3716,619	4123,631	3259,256	Bin Neg
	Chi-squared statistic	254,0406	331,409	86,70452	Bin Neg
Chromatin	Aikake's Information Criterion	3165,156	3067,771	2960,557	Bin Neg
	Bayesian Information Criterion	3174,209	3072,298	2969,610	Bin Neg
	Chi-squared statistic	840,2256	9775,434	665,8935	Bin Neg
Nucleoli	Aikake's Information Criterion	3465,746	3532,954	2974,323	Bin Neg
	Bayesian Information Criterion	3474,799	3537,481	2983,376	Bin Neg
	Chi-squared statistic	1390,681	2923,013	496,4243	Bin Neg
Mitoses	Aikake's Information Criterion	2692,113	2244,313	2177,956	Bin Neg
	Bayesian Information Criterion	2701,166	2248,839	2187,008	Bin Neg

En la Tabla 3.3 se puede observar como la mayoría de los atributos son representados por la distribución de probabilidad Binomial negativa, a excepción de Nuclei, pero que a pesar de esto, su AIC y BIC apuntan a que es buena idea considerar a la Binomial negativa igualmente. Esta tabla nos sirve como preliminar para entender el comportamiento de las variables.

En la Tabla 3.5 se muestra la correlación de Spearman entre cada combinación de par de atributos con lo que se obtiene una matriz triangular y el tipo de clasificación de correlación al cual pertenecen según la Tabla 3.4. Lógicamente todas las correlaciones de los atributos consigo mismos serán perfectas. Se utiliza esta correlación debido a que la naturaleza de las variables es cualitativa ordinal.

Tabla 3.4: Índices de correlación

Tipo	Nula	Muy Baja	Baja	Moderada	Alta	Muy Alta	Perfecta
Rango	$[0]$	$]0-0{,}2[$	$[0{,}2-0{,}4[$	$[0{,}4-0{,}6[$	$[0{,}6-0{,}8[$	$[0{,}8-1[$	$[1]$
Abreviación	**N**	**B-**	**B**	**M**	**A**	**A+**	**P**

Los Atributos en general presentan correlaciones altas entre sí, es considerable la correlación entre shape y size, quien nos indica que ambas ocurren en las mismas proporciones. Por otra parte, la variable clump y mitoses presentan correlaciones moderadas para casi con todas las variables, en algunos casos llega a ser baja, lo que nos indica que este atributo puede no ser un factor decisivo al nivel que lo son los demás.
Las correlaciones altas nos indican que a medida que ocurre una, ocurre la otra y en la misma dirección, vale decir, si una aumenta la otra también. por otra parte a medida que la correlación disminuye se puede afirmar con menor certeza la ocurrencia de una con la otra.

Tabla 3.5: Correlación de Spearman entre combinaciones de variables

Variable / Variable	clump	size	shape	adhesion	epithelial	nuclei	chromatin	nucleoli	mitoses
clump	1 P	0,6636 A	0,6666 A	0,5438 M	0,5867 M	0,5907 M	0,5338 M	0,5663 M	0,4212 B
size	–	1 P	0,8949 A+	0,7454 A	0,7928 A	0,7695 A	0,7205 A	0,7525 A	0,5126 M
shape	–	–	1 P	0,7187 A	0,7650 A	0,7529 A	0,6948 A	0,7244 A	0,4784 M
adhesion	–	–	–	1 P	0,6651 A	0,6967 A	0,6287 A	0,6364 A	0,4472 M
epithelial	–	–	–	–	1 P	0,6945 A	0,6450 A	0,7106 A	0,4832 M
nuclei	–	–	–	–	–	1 P	0,6789 A	0,6597 A	0,4743 M
chromatin	–	–	–	–	–	–	1 P	0,6620 A	0,3906 B
nucleoli	–	–	–	–	–	–	–	1 P	0,5102 M
mitoses	–	–	–	–	–	–	–	–	1 P

Por último en la figura 3-4 se puede observar el diagrama de cajas para los atributos en donde podemos analizar cuartiles de manera gráfica, de lo que se desprende que la mayoría de las instancias corresponden a niveles cercanos a la normalidad. Por otra parte, en los atributos chromatin, nuclei y clump estas instancias tienden a acumularse en sectores más alejados a la normalidad en comparación con el resto de sus compañeros. Caso distinto es en el de mitoses, quien tiene un comportamiento radicalmente distinto, del cual se puede desprender que las instancias que no se encuentren en niveles muy cercanos a la normalidad probablemente pertenezcan a objetos que sean de la clase cancerígena.

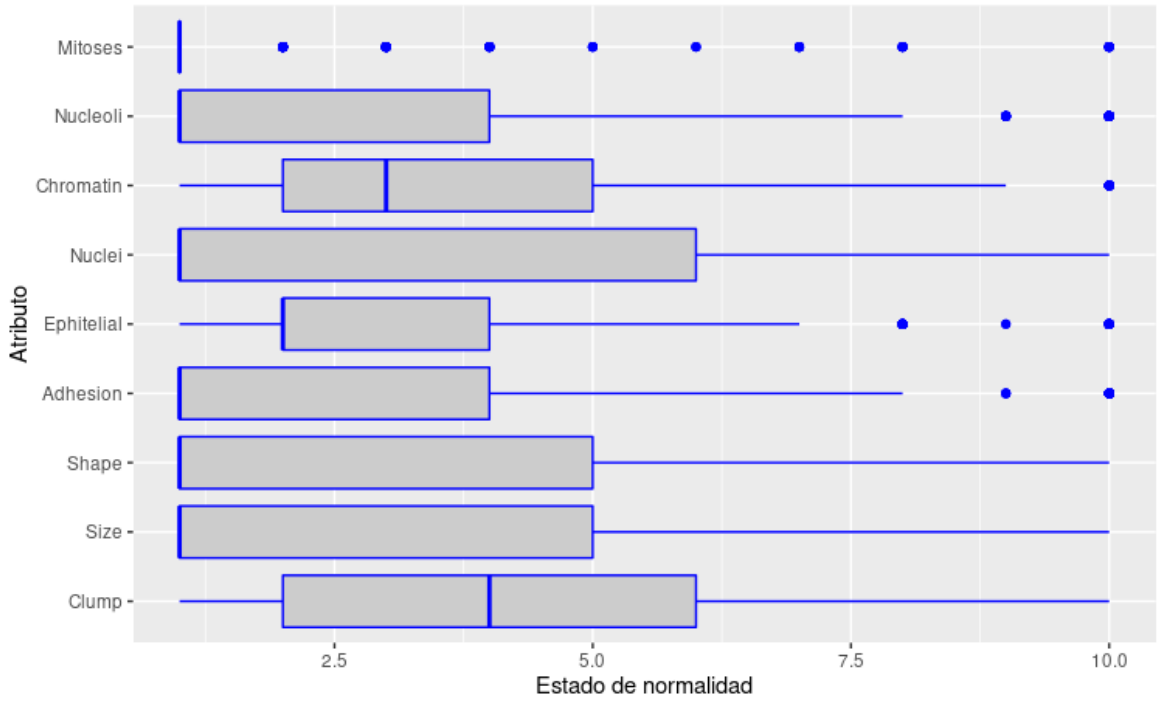

Figura 3-4: Boxplot

CAPÍTULO 4. CONCLUSIONES

A lo largo del presente informe se cumplieron los objetivos planteados, realizando un análisis estadístico e inferencial en base a la muestra de la base de datos estudiada. Se utilizó el software R como herramienta vital de apoyo para la abstracción de los mecanismos de cálculo y tabulación, que permitió abrir una nueva ventana para el aprendizaje de estadísticas, de manera que la visualización de indicadores estadísticos fue utilizable e intuitiva. Aunque se requirió de material adicional para enfrentar algunos problemas y así poder implementar funciones, no fue de gran dificultad hacer analogía con los lenguajes de programación aprendidos a lo largo de la carrera.

Se obtuvieron distintos estadísticos con el fin de definir los dominios de las variables, entender la coherencia de estas y cómo se comportaban a través de su rango definido; así como estadísticos de medidas de centralización para entender la distribución de la muestra y facilitar cálculos posteriores complejos; también las respectivas medidas de dispersión respecto a las medidas centrales, para entender la homogeneidad de los datos y por contraparte, su variación; y por último, entender las relaciones existentes entre variables, lo que nos permitió ver si tienen una relación perfecta, muy alta,alta, moderada, baja, muy baja o nula, con el fin de saber qué variable puede determinar en qué medida a cual.

Cabe destacar que a medida que los grados de clasificación varían de normal a anormal en cada atributo, las instancias de los primeros tres grados de clasificación del estado de dicho atributo tienen mayor probabilidad de pertenecer a objetos no cancerígenos, lo que concuerda con el mayor porcentaje de la muestra (65 %) que representa a tumores benignos.

Por otro lado, los atributos nuclei, clump y size en un estado anormal, serían las principales características que diferenciarían a un tumor maligno de uno benigno, además que el atributo size tiene una alta correlación con shape lo cual significa que si una de las dos se ve afectada la otra en algún momento también será afectada. Pero sería bueno realizar un análisis de componentes para una mejor precisión en las conclusiones de este análisis.

Por último queda la incertidumbre sobre las distribuciones obtenidas, ya que si se hubiese abordado este tema con muestras, utilizando a la base de dato como población, sería posible confiar en el teorema del límite central y así utilizar la distribución normal como representante, con el fin de utilizar métodos que la necesitan como requisito, como anova.

CAPÍTULO 5. BIBLIOGRAFÍA

[1] Abelardo Montesinos López. *ESTUDIO DEL AIC Y BIC EN LA SELECCIÓN DE MODELOS DE VIDA CON DATOS CENSURADOS*. 2011-08. URL: http://probayestadistica.cimat.mx/sites/default/files/PDFs/TE414MontesinosLopez.pdf (visitado 04-09-2016).

[2] MedinePlus. *Cancer*. 2016-08-23. URL: https://medlineplus.gov/spanish/ency/article/001289.htm (visitado 02-09-2016).

[3] Deepa Rao y Sujuan Zhao. *Prediction of Breast Cancer*. 2012-05-03. URL: https://gsm672.wikispaces.com/Prediction+of+Breast+cancer (visitado 01-09-2016).

[4] Wikipedia. *Cancer de Mama*. 2016-08-31. URL: https://es.wikipedia.org/wiki/C%C3%A1ncer_de_mama (visitado 31-08-2016).

[5] Wikipedia. *CSV*. 2016-08-10. URL: https://es.wikipedia.org/wiki/CSV (visitado 31-08-2016).

[6] William H. Wolberg. *Breast Cancer Wisconsin (Diagnostic) Data Set*. 1995-11-01. URL: http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29 (visitado 02-09-2016).

[7] William H. Wolberg y O.L. Mangasarian. «Multisurface method of pattern separation for medical diagnosis applied to breast cytology». En: *Proceedings of the National Academy of Sciences* 87 (1990), pp. 9193-9196.

CAPÍTULO 6. APENDICE

```
library("modeest")
library("stats")
library("ggplot2")
library("MASS")
library("survival")
library("fitdistrplus")
library("gridExtra")

#coeficiente de variacion
coefVar<-function(des,med){
  coef<-(des/(abs(med)))
  return(coef)
}

#importacion de datos
datos<-read.csv2("breast-cancer-wisconsin.csv",header=TRUE,sep = ",")

#medidas de centralizacion
media<-c()
moda<-c()
mediana<-c()

#medidas de variacion
varianza<-c()
cVariacion<-c()

#Obtencion de datos
for(i in 1:9){
    media<-c(media,mean(datos[,i]))
    moda<-c(moda,mfv(datos[,i]))
    mediana<-c(mediana,median(datos[,i]))
    varianza<-c(varianza,var(datos[,i]))

}
```

```
#obtencion de coeficientes de variacion
cVariacion<-coefVar(sqrt(varianza),media)

#resumen
medidas<-cbind(moda,mediana,media,varianza,cVariacion)

#taba de frecuencias
tablas<-c()
tablas<-rbind(tablas,table(datos$clump),table(datos$size),table(datos$shape),
    table(datos$adhesion),table(datos$epithelial),table(datos$nuclei),
    table(datos$chromatin),table(datos$nucleoli),table(datos$mitoses))
tablaClass<-table(datos[,10])

#correlaciones de spearman ###################
varS12<-cor(datos$clump,datos$size,method="spearman")
varS13<-cor(datos$clump,datos$shape,method="spearman")
varS14<-cor(datos$clump,datos$adhesion,method="spearman")
varS15<-cor(datos$clump,datos$epithelial,method="spearman")
varS16<-cor(datos$clump,datos$nuclei,method="spearman")
varS17<-cor(datos$clump,datos$chromatin,method="spearman")
varS18<-cor(datos$clump,datos$nucleoli,method="spearman")
varS19<-cor(datos$clump,datos$mitoses,method="spearman")

varS23<-cor(datos$size,datos$shape,method="spearman")
varS24<-cor(datos$size,datos$adhesion,method="spearman")
varS25<-cor(datos$size,datos$epithelial,method="spearman")
varS26<-cor(datos$size,datos$nuclei,method="spearman")
varS27<-cor(datos$size,datos$chromatin,method="spearman")
varS28<-cor(datos$size,datos$nucleoli,method="spearman")
varS29<-cor(datos$size,datos$mitoses,method="spearman")

varS34<-cor(datos$shape,datos$adhesion,method="spearman")
varS35<-cor(datos$shape,datos$epithelial,method="spearman")
varS36<-cor(datos$shape,datos$nuclei,method="spearman")
varS37<-cor(datos$shape,datos$chromatin,method="spearman")
varS38<-cor(datos$shape,datos$nucleoli,method="spearman")
varS39<-cor(datos$shape,datos$mitoses,method="spearman")
```

```
varS45<-cor(datos$adhesion,datos$epithelial,method="spearman")
varS46<-cor(datos$adhesion,datos$nuclei,method="spearman")
varS47<-cor(datos$adhesion,datos$chromatin,method="spearman")
varS48<-cor(datos$adhesion,datos$nucleoli,method="spearman")
varS49<-cor(datos$adhesion,datos$mitoses,method="spearman")

varS56<-cor(datos$epithelial,datos$nuclei,method="spearman")
varS57<-cor(datos$epithelial,datos$chromatin,method="spearman")
varS58<-cor(datos$epithelial,datos$nucleoli,method="spearman")
varS59<-cor(datos$epithelial,datos$mitoses,method="spearman")

varS67<-cor(datos$nuclei,datos$chromatin,method="spearman")
varS68<-cor(datos$nuclei,datos$nucleoli,method="spearman")
varS69<-cor(datos$nuclei,datos$mitoses,method="spearman")

varS78<-cor(datos$chromatin,datos$nucleoli,method="spearman")
varS79<-cor(datos$chromatin,datos$mitoses,method="spearman")

varS89<-cor(datos$nucleoli,datos$mitoses,method="spearman")

corS<-c(varS12,varS13,varS14,varS15,varS16,varS17,varS18,varS19,varS23,
    varS24,varS25,varS26,varS27,varS28,varS29,varS34,varS35,varS36,
    varS37,varS38,varS39,varS45,varS46,varS47,varS48,varS49,varS56,
    varS57,varS58,varS59,varS67,varS68,varS69,varS78,varS79,varS89)
#####################################

#analisis de posibles  distribuciones
descdist(datos$clump,discrete = TRUE)
descdist(datos$size,discrete = TRUE)
descdist(datos$shape,discrete = TRUE)
descdist(datos$adhesion,discrete = TRUE)
descdist(datos$epithelial,discrete = TRUE)
descdist(datos$nuclei,discrete = TRUE)
descdist(datos$chromatin,discrete = TRUE)
descdist(datos$nucleoli,discrete = TRUE)
descdist(datos$mitoses,discrete = TRUE)
```

```
#Ajuste de distribuciones normal, poisson y negativa binomial.
clump.n<-fitdist(datos$clump,"norm")
clump.p<-fitdist(datos$clump,"pois")
clump.nb<-fitdist(datos$clump,"nbinom")
cdfcomp(list(clump.n,clump.p,clump.nb),
    legendtext =c("Normal","Poison","Binomial negativa") )
gofstat(list(clump.n,clump.p,clump.nb),
    fitnames =c("Normal","Poison","Binomial negativa"),discrete = TRUE )

size.n<-fitdist(datos$size,"norm")
size.p<-fitdist(datos$size,"pois")
size.nb<-fitdist(datos$size,"nbinom")
cdfcomp(list(size.n,size.p,size.nb),
    legendtext =c("Normal","Poison","Binomial negativa") )
gofstat(list(size.n,size.p,size.nb),
    fitnames =c("Normal","Poison","Binomial negativa"),discrete = TRUE )

shape.n<-fitdist(datos$shape,"norm")
shape.p<-fitdist(datos$shape,"pois")
shape.nb<-fitdist(datos$shape,"nbinom")
cdfcomp(list(shape.n,shape.p,shape.nb),
    legendtext =c("Normal","Poison","Binomial negativa") )
gofstat(list(shape.n,shape.p,shape.nb),
    fitnames =c("Normal","Poison","Binomial negativa"),discrete = TRUE )

adhesion.n<-fitdist(datos$adhesion,"norm")
adhesion.p<-fitdist(datos$adhesion,"pois")
adhesion.nb<-fitdist(datos$adhesion,"nbinom")
cdfcomp(list(adhesion.n,adhesion.p,adhesion.nb),
    legendtext =c("Normal","Poison","Binomial negativa") )
gofstat(list(adhesion.n,adhesion.p,adhesion.nb),
    fitnames =c("Normal","Poison","Binomial negativa"),discrete = TRUE )

epithelial.n<-fitdist(datos$epithelial,"norm")
epithelial.p<-fitdist(datos$epithelial,"pois")
epithelial.nb<-fitdist(datos$epithelial,"nbinom")
```

```
cdfcomp(list(epithelial.n,epithelial.p,epithelial.nb),
    legendtext =c("Normal","Poison","Binomial negativa") )
gofstat(list(epithelial.n,epithelial.p,epithelial.nb),
    fitnames =c("Normal","Poison","Binomial negativa"),discrete = TRUE )

nuclei.n<-fitdist(datos$nuclei,"norm")
nuclei.p<-fitdist(datos$nuclei,"pois")
nuclei.nb<-fitdist(datos$nuclei,"nbinom")
cdfcomp(list(nuclei.n,nuclei.p,nuclei.nb),
    legendtext =c("Normal","Poison","Binomial negativa") )
gofstat(list(nuclei.n,nuclei.p,nuclei.nb),
    fitnames =c("Normal","Poison","Binomial negativa"),discrete = TRUE )

chromatin.n<-fitdist(datos$chromatin,"norm")
chromatin.p<-fitdist(datos$chromatin,"pois")
chromatin.nb<-fitdist(datos$chromatin,"nbinom")
cdfcomp(list(chromatin.n,chromatin.p,chromatin.nb),
 legendtext =c("Normal","Poison","Binomial negativa") )
gofstat(list(chromatin.n,chromatin.p,chromatin.nb),
 fitnames =c("Normal","Poison","Binomial negativa"),discrete = TRUE )

nucleoli.n<-fitdist(datos$nucleoli,"norm")
nucleoli.p<-fitdist(datos$nucleoli,"pois")
nucleoli.nb<-fitdist(datos$nucleoli,"nbinom")
cdfcomp(list(nucleoli.n,nucleoli.p,nucleoli.nb),
    legendtext =c("Normal","Poison","Binomial negativa") )
gofstat(list(nucleoli.n,nucleoli.p,nucleoli.nb),
    fitnames =c("Normal","Poison","Binomial negativa"),discrete = TRUE )

mitoses.n<-fitdist(datos$mitoses,"norm")
mitoses.p<-fitdist(datos$mitoses,"pois")
mitoses.nb<-fitdist(datos$mitoses,"nbinom")
cdfcomp(list(mitoses.n,mitoses.p,mitoses.nb),
    legendtext =c("Normal","Poison","Binomial negativa") )
gofstat(list(mitoses.n,mitoses.p,mitoses.nb),
    fitnames =c("Normal","Poison","Binomial negativa"),discrete = TRUE )
####################################
```

```
#Diagrama de caja
clump<-cbind(datos[,1],1)
size<-cbind(datos[,2],2)
shape<-cbind(datos[,3],3)
adhesion<-cbind(datos[,4],4)
ephitelial<-cbind(datos[,5],5)
nuclei<-cbind(datos[,6],6)
chromatin<-cbind(datos[,7],7)
nucleoli<-cbind(datos[,8],8)
mitoses<-cbind(datos[,9],9)

dt<-data.frame(rbind(clump,size,shape,adhesion,ephitelial,nuclei,
    chromatin,nucleoli,mitoses))
dt$X2 = factor(dt$X2,labels = c("Clump", "Size","Shape",
    "Adhesion","Ephitelial","Nuclei",
    "Chromatin","Nucleoli","Mitoses"))

ggplot(dt,aes(x=dt$X2,y=dt$X1)) +
  geom_boxplot(fill = "grey80", colour = "blue") +
  scale_x_discrete() + xlab("Atributo") +
  ylab("Estado de normalidad")+
  coord_flip()
###################################

#Diagramas de proporcion de atributos respecto a Clase.
datos$class<-as.factor(datos$class)
p1<-ggplot(datos,aes(datos$clump,fill=datos$class))
    +geom_bar()+theme_bw()+scale_fill_brewer(palette = "Set1")
p2<-ggplot(datos,aes(datos$size,fill=datos$class))
    +geom_bar()+theme_bw()+scale_fill_brewer(palette = "Set1")
p3<-ggplot(datos,aes(datos$shape,fill=datos$class))
    +geom_bar()+theme_bw()+scale_fill_brewer(palette = "Set1")
p4<-ggplot(datos,aes(datos$adhesion,fill=datos$class))
    +geom_bar()+theme_bw()+scale_fill_brewer(palette = "Set1")
p5<-ggplot(datos,aes(datos$epithelial,fill=datos$class))
    +geom_bar()+theme_bw()+scale_fill_brewer(palette = "Set1")
```

```
p6<-ggplot(datos,aes(datos$nuclei,fill=datos$class))
    +geom_bar()+theme_bw()+scale_fill_brewer(palette = "Set1")
p7<-ggplot(datos,aes(datos$chromatin,fill=datos$class))
    +geom_bar()+theme_bw()+scale_fill_brewer(palette = "Set1")
p8<-ggplot(datos,aes(datos$nucleoli,fill=datos$class))
    +geom_bar()+theme_bw()+scale_fill_brewer(palette = "Set1")
p9<-ggplot(datos,aes(datos$mitoses,fill=datos$class))
    +geom_bar()+theme_bw()+scale_fill_brewer(palette = "Set1")
grid.arrange(p1,p2,p3,p4,p5,p6,p7,p8,p9,ncol=2)
####################################
```