**CAPÍTULO II**

1. **MARCO TEÓRICO CONCEPTUAL DEL PROCESO DE SENSIBILIDAD Y LOS MÉTODOS ESTADÍSTICOS**

**Introducción**

En este capítulo se indica la descripción del proceso de sensibilidad para la determinación del efecto de los productos químicos sobre *Mycosphaerella f.*, en el cual se explica la selección y recolección de los datos, procesamiento en laboratorio y resultados; la población objetivo y la descripción de las variables a utilizar.

En el marco teórico se indica la metodología estadística necesaria para realizar el presente estudio, donde se detallan los objetivos, características e interpretación de cada método utilizado.

Se tratará como punto inicial la estadística descriptiva, continuando con la estadística inferencial para ello se realizan pruebas de hipótesis, uso de modelos lineales como la regresión simple y finalmente el uso de modelos no lineales de clasificación la regresión Logit y la regresión Probit.

Es importante definir el concepto de minería de datos que es parte de la metodología utilizada para el desarrollo de ciertos modelos estadísticos usados; y finalmente, qué tipos de software fueron utilizados durante la construcción de la aplicación.

* 1. Descripción del proceso de sensibilidad.

El proceso de sensibilidad proporcionado por el CIBE, comienza con las entrevistas sobre los requerimientos de necesidad del cliente, donde se proveen los productos a utilizar y dosis a suministrar; para el estudio se seleccionó el fungicida sistémico Azoxistrobina (Bankit).

La siguiente etapa del proceso es la selección y recolección de manera dirigida de hojas afectadas con Sigatoka negra; luego en laboratorio las muestras pasan al proceso de incubación y colocadas en cajas petri con medios de cultivo por 2 horas; después de 24 horas bajo el lente del microscopio se mide el tamaño de los tubos germinativos para evaluar su crecimiento, es necesario mencionar que en este proceso también se mide el tamaño de la colonia de *Mycosphaerella f.* como parte del método utilizado y que es de igual importancia en los análisis que se realizan en el laboratorio.

Los datos obtenidos son almacenados en hojas electrónicas, previo al análisis estadístico, y se procede a realizar el cálculo de la Dosis Letal Media mediante técnicas estadísticas, que se ampliará la explicación de dichas técnicas mediante el uso de la herramienta informática más adelante.

Posteriormente se prepara el informe, el cual es revisado y entregado al cliente para su satisfacción, y finalmente se respaldan los resultados en formato digital.[[1]](#footnote-1)

* 1. Población Objetivo.

Se seleccionó y recopiló un número de hojas infectadas según la escala Stover modificada con grado de infección 4 y 5 *(Gauhl, 1989)*, dichas hojas fueron recolectadas de manera dirigida de la finca *“El Paraíso”*, Santa Ana – Sto. Domingo de los Colorados en el 2005. El número de observaciones es de 150 unidades, las cuales fueron obtenidas en los laboratorios del CIBE, esto sucede durante el proceso de sensibilidad descrito en el literal 2.1.

Es necesario mencionar que el análisis de sensibilidad dependerá generalmente de los requerimientos del cliente, en cuanto a producto, dosis y muestras.

* 1. Descripción de las variables.

En los ensayos se utilizan tanto productos orgánicos como productos químicos ó convencionales, cabe recalcar que se utilizan dos tipos de medida para las dosis suministradas, tal como el porcentaje (%) del producto utilizado, por ejemplo: dosis al 0.5%, 1%, 10%; y partes por millón (ppm) del producto, por ejemplo 0.001 ppm, 0.00001 ppm; con esto se obtendrán los diferentes tratamientos ó concentraciones para el análisis.

Las variables de estudio serán el crecimiento del tubo germinativo y el diámetro de la colonia de *Mycosphaerella f.*, las que tienen como escala de medición micras y milímetros respectivamente; estas variables de origen o iníciales son necesarias para determinar la variable principal para el análisis y es en la que se fundamenta nuestro estudio, la variable porcentaje de inhibición.

El porcentaje de inhibición es obtenido a través de la siguiente fórmula:

Donde es el promedio de la medida de control ó dosis 0; es igual a la medida de la unidad de observaciones a la que se le suministró cierta dosis del producto químico en estudio.[[2]](#footnote-2)

* 1. **Estadística Descriptiva**

El análisis descriptivo de los datos, tiene como objetivo obtener una visión general del comportamiento de las variables, para ello se obtienen medidas de tendencia central: media y moda; medidas de dispersión: varianza, desviación estándar, rango y cuartiles; distribución de los datos: sesgos y curtosis, que indica que tan distantes se encuentran los datos respecto al promedio; y percentiles que es la división de los datos ordenados y divididos en cien partes, los que pueden ser representados gráficamente mediante la gráfica denominada Ojiva y su presentación más común es el diagrama de cajas, donde se observan los percentiles (P25, P50, P75).



 FIGURA 2.1 DIAGRAMA DE CAJA.

 FUENTE: N.M. DOWNIE Y R.W. HEATH (1986)

El análisis descriptivo se utilizó para conocer la distribución de los datos, la existencia de valores atípicos, sesgos, que pudieran dificultar el análisis inferencial y si es necesario realizar alguna transformación matemática.[[3]](#footnote-3)

* 1. **Análisis de varianza para un solo factor.**

El análisis de varianza es una de las técnicas usada en el diseño experimental, la cual ayuda a probar si existen diferencias significativas de al menos uno de los promedios de la variable de estudio, obtenido de los tratamientos de diferentes poblaciones.

Diseño experimental es una técnica estadística, donde la primera fase es planificar un ensayo de tal manera que este provea de la información pertinente sobre el problema investigado. Existen varios tipos de diseños tales como el completamente al azar, de bloques completos al azar, experimentos factoriales, entre otros, los cuales son utilizados de acuerdo a los objetivos de cada experimento.

Para el propósito de la investigación, el análisis y resultados estarán fundamentados en un diseño completamente aleatorio DCA, el mismo que se examinará a través del análisis de varianza (ANOVA) de un solo factor, donde sus diferentes niveles son las dosis probadas, y se tratará de explicar la variable cuantitativa a través de una variable cualitativa.

La Tabla 2.1 muestra la matriz de observaciones, donde se tienen ***a*** tratamientos diferentes de un misma población que quieren compararse, las *n* observaciones tomadas del experimento en cada tratamiento, constituyendo una matriz de datos de *axn*, donde *yij* es la observación j-ésima tomada del i-ésimo tratamiento.[[4]](#footnote-4)

|  |
| --- |
| **Tabla 2.1** |
| **Observaciones de un análisis de varianza de un solo factor** |
| Tratamiento (nivel) | Observaciones |
| 1 | y11 | y12 | … | y1n |
| 2 | y21 | y22 | … | y2n |
| … | … | … | … | … |
| a | ya1 | ya2 | … | Yan |

FUENTE: MONTGOMERY D. (2003)

En este contexto se describirá mejor las observaciones del experimento mediante el modelo siguiente

 **(2.1)**

Donde es la media total de experimento, es el efecto del i-ésimo tratamiento y es el error aleatorio por cada observación. Siendo el supuesto para los errores, estos deben seguir una distribución normal con media igual a cero y varianza constante y la suma de los efectos por tratamiento debe ser igual a cero

Se define la hipótesis en el Tabla 2.2, para probar la igualdad de las ***a*** medias de los tratamientos, donde el contraste de hipótesis es:

|  |
| --- |
| **Tabla 2.2****Contraste de Hipótesis de diferencias de medias** |
| Vs. |

FUENTE: MONTGOMERY D. (2003)

Se procede a la partición de la variabilidad total, que ayudará al análisis correcto de los datos en cada tratamiento, la Tabla 2.3 muestra los componentes de la varianza total denotada por *SST* suma cuadrática total, con *N-1* grados de libertad, siendo esta la misma medida de variabilidad global de los datos; la suma cuadrática de los tratamientos (*SStr*), que mide la variabilidad que existe entre los tratamientos, con *a-1* grados de libertad; y la suma cuadrática del error (*SSe*), que es una medida de variabilidad dentro de los tratamientos, que se atribuye al error aleatorio, con *N-a* grados de libertad; formando la ecuación descrita a continuación

 *SST = SStr + SSe* **(2.2)**.

|  |
| --- |
| **Tabla 2.3** |
| **Tabla de análisis de varianza para un solo tratamiento** |
| **Fuente de variación** | **Suma de los Cuadrados** | **Grados de libertad** | **Medias de los Cuadrados** | **F0** |
| Entre los tratamientos |  | a-1 |  |  |
| Dentro de los tratamientos (Error) |  | N-a |  |  |
| Total |  | N-1 |  |  |

\* ; ,i =1,2,..,a

 ;

FUENTE: MONTGOMERY D. (2003)

Bajo los supuestos de normalización e independencia de los errores () y las observaciones (), se usa el Teorema de Cochran en el Anexo 3, que establece que la Suma de los cuadrados de los tratamientos y de los errores son variables aleatorias Ji-cuadrada con una distribución independiente, por lo tanto se postula el estadístico de prueba sobre la hipótesis propuesta en Cuadro 2.1, que se distribuye con *F*, con *a-1* y *N-a* grados de libertad, donde con (1- )% de confianza se rechaza la hipótesis nula a favor de la alternativa, si > o tomando el *P-Valor* como criterio de decisión. Al rechazar la hipótesis nula, se asegura que existe evidencia estadística para afirmar que al menos uno de los promedios es diferente y que por lo tanto existe un efecto del factor, diferente de cero, en la variable de estudio.

* 1. **Regresión Lineal Simple.**

Uno de los modelos de regresión a desarrollar es el de regresión lineal simple, el cual tiene como objetivo determinar un modelo lineal que se ajuste a los datos de este estudio[[5]](#footnote-5), el método busca la relación funcional entre dos variables, una variable llamada independiente ó explicativa y otra variable llamada dependiente o de respuesta, el cual se representa por la siguiente ecuación de predicción:

 **(2.3)**

 es el valor donde la línea de regresión se intercepta con el eje *Y,* es el coeficiente de regresión ó la pendiente de la línea recta. Para estimar el modelo y los valores de la variable dependiente, se obtienen los estimadores de los parámetr**os** ;y al lado derecho de la ecuación 2.3 se le debe agregar que es el error aleatorio asociado a la medición de *X*.

Se supone que el error aleatorio tiene una distribución normal con media cero y varianza constante y la relación entre las variables es lineal.

Mediante el uso del método de los mínimos cuadrados, se procede a calcular los y que son los estimadores de y respectivamente, partiendo de una muestra de *n* pares de observaciones de la variable independiente y dependiente tales como , para minimizar el error se calcula la suma cuadrada de los errores mediante la ecuación siguiente:

 **(2.4)**

Luego se deriva con respecto a los coeficientes de regresión e igualándolos a cero se obtienen las ecuaciones finales para encontrar dichos coeficientes.

 **(2.5)**

 **(2.6)**

Donde y que son los promedios de las variables independientes y dependientes. Para el presente caso de estudio, la variable independiente ***X***  es el logaritmo natural de la dosis (Ln(X)) y la dependiente ***Y***  es el porcentaje de inhibición, una vez estimados estos valores se los reemplaza en la ecuación de predicción expuesta en (2.3) la que nos servirá como modelo en búsqueda de la Dosis Letal, para el cálculo de la DL50 se iguala dicha ecuación para ***Y*** = 50, y despejando la variable ***X*** se obtendrá el valor de la dosis letal media.

* 1. **Regresión LOGIT y PROBIT.**

Es un análisis mediante el uso de modelos no lineales en la investigación, que tiene como objetivo estimar las relaciones de una o más variables respuestas en términos de otras a través de una función no lineal que haga posible un mejor ajuste de los datos de estudio y que haga posible predecir o explicar las decisiones e influencias tomadas sobre un conjunto de variables. Los modelos de regresión a usarse para efectos de la investigación son Logit y Probit, con los cuales se determinará si tienen o no las variables independientes *X1, X2,..., Xn*, un efecto sobre la variable dependiente, descrita por una respuesta a dos posibles eventos *Y1* y *Y2*, formando la relación funcional entre *Y* y *X*; tal como *Y = ( X1, X2,..., Xn).[[6]](#footnote-6)*

* + 1. **Tipos y codificación de Variables de los modelos.**

Se definen los tipos de variables que tomarán los modelos no lineales mencionados:

***Variable Dependiente:*** es aquella variable binaria o dicotómica que solo puede tomar dos posibles valores (Éxito o fracaso, ej: Si–No, 0-1, Verdadero-Falso). Este tipo de variable es un subconjunto de las llamadas variables categóricas o cualitativas.

***Variables Independientes:*** pueden ser de los tipos a continuación:

***Binaria*** es aquella que solo puede tomar dos posibles valores, 1 si ocurre y 0 caso contrario.

***Variable Categórica Nominal***: aquella variable que puede tomar varios valores que representan las categorías sin tener algún orden específico (1=Hombre, 2=Mujer).

***Variable Categórica Ordinal***: aquella variable que puede tomar varios valores que representan las categorías con un orden específico (1=Bajo, 2=Medio, 3=Alto).

***Dummy***: Si una variable categórica nominal u ordinal que puede tomar más de dos valores, y se la puede codificar en varias variables binarias o dicotómicas, esta puede ser llamada variable indicadora dummy.

***Cuantitativa***: aquella variable que puede tomar un rango numérico de valores (-, +), los cuales puede ser discreta (numerable) o continua (x ∈ R).

* + 1. **Modelo Logit**

Se procede a construir el modelo de regresión logística donde se tiene una variable dependiente dicotómica o predictora de manera
 ,

y un conjunto de variables independientes *X1, X2,..., Xn*. Se plantea el siguiente modelo

 **(2.7)**

Se estiman los coeficientes del modelo por el procedimiento de mínimos cuadrados, como un modelo de regresión lineal, sin embargo esto llevaría a resultados ilógicos, debido a que los resultados estarán fuera del rango [0–1], causando una contradicción con la variable dependiente planteada.[[7]](#footnote-7)

Se procede a construir el modelo logístico que establece una relación entre la probabilidad de que ocurra un evento, dado un conjunto de variables explicativas, de la siguiente forma

 **(2.8)**

Ahora se tiene un modelo que puede tener cualquier valor, que mediante transformaciones logarítmicas se obtiene la ecuación equivalente a

 **(2.9)**

* + 1. **Modelo Probit**

El modelo de regresión Probit es otro método para explicar la variable dependiente dicotómica, a través de las variables independientes.[[8]](#footnote-8)

Tomando de la ecuación 2.8se explica el modelo usando la distribución normal estándar, que se describe a continuación

 **(2.10)**

Donde los coeficientes reflejan el impacto que tiene la variable independiente sobre la probabilidad, coeficientes que son estimados para ambos modelos mediante el método de Máxima Verosimilitud, que se explica con detalle en el literal 2.7.6.1.

* + 1. **Coeficientes de riesgo**

***Riesgo* (probabilidad)**: se define como una medida de probabilidad de un evento, mediante el cociente del número de veces (casos) que ocurre un evento frente al número total de casos

***Odds* (oportunidad)**: se define como una medida de ocurrencia de un evento, mediante el cociente del número de casos que ocurre un evento frente al número de casos que no ocurre, que es una manera de decir lo mismo que con probabilidades.

 **(2.11)**

Siendo *p* la probabilidad del suceso, mientras más probable se produzca el evento frente a que no se produzca, dicho cociente fue el principio para construir los modelos planteados.

***Odds Ratio* (oportunidad relativa)**: se define como el cociente del Odds correspondiente al evento bajo cierta condición entre el Odds del mismo evento bajo otra condición.

 **(2.12)**

Por ejemplo, la probabilidad de aparición de hipertensión en una persona cuando se consume mucha sal (odds1) respecto a la probabilidad del mismo evento cuando consume poca sal (odds2).

Los Odds ratio resultan de gran utilidad para evaluar la influencia que cada variable independiente tiene sobre la variable de respuesta, donde un Odds ratio igual a uno indica igual probabilidad bajo un evento con diferentes condiciones, mayor a uno indica un aumento en la probabilidad del evento y menor a uno implica disminución sobre dicha probabilidad.[[9]](#footnote-9)

* + 1. **Estimación de coeficientes**

Se procede a estimar los coeficientes del modelo, denotados por el vector , partiendo de una muestra aleatoria de tamaño *n* dada por , donde es el valor de las variables independientes e es el valor observado de y en el i-ésimo elemento de la muestra. Utilizando el hecho de que y| tiene una distribución binomial , ya que la variable dependiente toma dos posibles valores éxito y fracaso, cuando el numero de éxitos en *n* repeticiones tiene un distribución binomial *B(n, p)*. Entonces la función de verosimilitud vendrá dada por

 **(2.13)**

Donde

* + 1. **Pruebas de Hipótesis.**

Las pruebas a continuación implican formular y probar una hipótesis estadística para determinar si las variables independientes del modelo se encuentran relacionadas con la variable dependiente, mediante una comparación de los valores observados respecto a los valores que se han predicho por los modelos utilizados.[[10]](#footnote-10)

* + - 1. **Criterio de Razón de Verosimilitudes.**

Se plantea el siguiente contraste de hipótesis en la Tabla 2.4 para determinar si las variables independientes influyen o aportan significativamente al modelo encontrado y si explican a la variable dependiente, bajo el criterio de *Razón de Verosimilitudes*

|  |
| --- |
| **Tabla 2.4****Contraste de Hipótesis para el criterio de Razón de Verosimilitudes** |
| Vs.Estadístico de Prueba es :  |

FUENTE: MONTGOMERY D. (2003)

Donde son los estimadores máximo verosímiles de los vectores de parámetros del modelo observado y el predictor.

El criterio de decisión será si grados de libertad, se concluye que al menos una variable aporta significativamente al modelo.

* + - 1. **Criterio de Hosmer-Lemeshow.**

Se procede a evaluar el modelo seleccionado, mediante el criterio de *Hosmer-Lemeshow*, que indica si el modelo obtenido se ajusta a los valores observados, se muestra el contraste de hipótesis en la Tabla 2.5.

|  |
| --- |
| **Tabla 2.5****Contraste de Hipótesis para el criterio de Hosmer-Lemeshow.** |
| Vs.Estadístico de Prueba es :  |

FUENTE: MONTGOMERY D. (2003)

El criterio de decisión será si , no se rechaza la hipótesis nula y se concluye que el modelo se ajusta a los datos observados.

* + - 1. **Criterio de Wald.**

Una vez detectada la existencia de un conjunto de variables explicativas que influyen en el modelo, se procede a identificar cuáles son estas variables, mediante *La Prueba de Wald***,** se muestra el contraste de hipótesis en la Tabla 2.6.

|  |
| --- |
| **Tabla 2.6****Contraste de Hipótesis para el criterio de Wald** |
| Vs.Estadístico de Prueba es :  |

FUENTE: MONTGOMERY D. (2003)

El criterio de decisión será si entonces se rechaza la hipótesis nula a favor de la alterna, concluyendo que la variable independiente correspondiente al coeficiente evaluado influye a la probabilidad del evento.

* 1. **Minería de datos**

Una de las tendencias a nivel de pequeñas, medianas y grandes empresas es diseñar y construir sus propias aplicaciones o sistemas informáticos para realizar actividades específicas, que ayudarán al correcto almacenamiento de la información, mejora de los procesos, optimización de los tiempos y confiabilidad de los resultados.

La minería de datos es un término utilizado para la exploración y análisis de los datos con el fin de encontrar patrones y reglas que expliquen el comportamiento de los datos. La minería de datos se encarga principalmente en la construcción de modelos, estos modelos no son más que algoritmos en los que se especifican entradas, procesos y salidas, que servirán para descubrir el conocimiento de los datos.

Muchas de las tareas a cargo de la minería de datos son la clasificación, estimación, predicción, agrupación, reglas de asociación y perfiles, los que son desarrollados en el campo de la ingeniería, economía, biología, modelos de negocios, etc. como por ejemplo el desarrollo de un modelo predictivo del brote de la Sigatoka Negra para las plantaciones de plátano al sur del lago de Maracaibo.[[11]](#footnote-11)

En fin, se pretende aclarar ciertos conceptos que se encuentran inmersos y hasta de manera oculta pero que son muy importantes para el desarrollo de la herramienta informática, la misma que permitirá acceder a la información que se encuentra almacenada en una base de datos, esta información será procesada por medio de algoritmos que ejecutan técnicas de aprendizaje sobre el 25% de los datos; y luego los algoritmos de regresión Logit y Probit, para la búsqueda de la dosis de fungicida que se aplicará a los cultivos.

* 1. **Software utilizado para el desarrollo de la herramienta.**

Las herramientas informáticas ó software requeridos para la ejecución de la aplicación fue Excel 2003 como la interfaz principal para el usuario y PostgreSql 8.2 que es la base de datos para el almacenamiento de la información; cabe recalcar que la base de datos utilizada está en la categoría de software libre, esto significa que no incurre en algún costo por compra o licencias anuales, además es un software completo que cuenta con herramientas y funciones ideales para el desarrollo de este proyecto.

1. Véase en bibliografía [24]. [↑](#footnote-ref-1)
2. Véase en bibliografía [24]. [↑](#footnote-ref-2)
3. Véase en bibliografía [8]. [↑](#footnote-ref-3)
4. Véase en bibliografía [18]. [↑](#footnote-ref-4)
5. Véase en bibliografía [18]. [↑](#footnote-ref-5)
6. Véase en bibliografía [4], [28] y [31]. [↑](#footnote-ref-6)
7. Véase en bibliografía [28] y [31] [↑](#footnote-ref-7)
8. Véase en bibliografía [28] y [31] [↑](#footnote-ref-8)
9. Véase en bibliografía [28] y [31] [↑](#footnote-ref-9)
10. Véase en bibliografía [28] y [31] [↑](#footnote-ref-10)
11. Véase en bibliografía [7] [↑](#footnote-ref-11)