Capítulo I

Introducción a la Estadística

Introducción

¿Qué es Estadística?

Estamos rodeados por información. Esto es cierto tanto en nuestra vida no profesional como en nuestra vida como científicos y biólogos. Hay en efecto tanta información para manejar, que muy a menudo se necesitan formas de resumirla, condensarla e interpretarla. El grupo de técnicas y teorías desarrolladas con este propósito se llaman colectivamente **estadística** y cada una de ellas por separado es llamada un **estadístico.**

Los biólogos (o científicos en general) estudian estadística porque uno de los componentes más importantes y fundamentales de este tipo de trabajo es la información que comúnmente llamamos **datos**. Se obtienen datos de casi todo, desde la simple lectura de un termómetro hasta el más complejo estudio de varios años sobre las interacciones a varios niveles de una comunidad. Sin una forma de resumir y analizar estos datos solo tendríamos listas de números sin mucho significado. La estadística se usa como ayuda para interpretar estas observaciones y sacar conclusiones interesantes desde el punto de vista biológico.

Clases de Estadística:

A través de este libro se hará referencia a dos clases de estadística: **Descriptiva** e **Inferencial**. Los nombres de estas dos clases se relacionan con los dos papeles principales de la estadística: descripción de datos e inferencia a partir de ellos. Sin embargo, se debe ver que esta distinción es arbitraria y que ambas pueden, en diferentes ocasiones, ser descriptivas o inferenciales.

Estadística Descriptiva:

Cubre las técnicas que se usan para resumir datos. Si existe algún problema para entender por qué se podría querer resumir datos, imagine que se tiene que describir a un director de tesis los resultados de 50.000 observaciones de comportamiento, ¡sin resumirlos!.

Hay dos maneras principales para resumir datos. La primera consiste en describir los datos numéricamente. Esto incluye rangos, medias y varianzas que proveen una forma fácil y concisa de resumir datos. Sin embargo, sólo dan una cantidad de información limitada acerca de ellos. La segunda y más efectiva manera de resumirlos es presentarlos en forma de dibujos. La representación visual puede mostrar concisamente la riqueza de información que se tenga. Los dibujos (gráficos, diagramas y mapas) son herramientas indispensables en el análisis de datos y deben ser casi siempre considerados como el primer paso a dicho análisis.

Hay muchas formas numéricas y visuales de resumir casi cualquier grupo de datos. Decidir cuál aplicar a un problema particular depende de la naturaleza de los datos así como de cierta cantidad de imaginación y creatividad.

Estadística Inferencial:

Estas son pruebas que permiten sacar conclusiones acerca de un grupo grande de organismos, en base a la información obtenida a partir de una muestra pequeña del total. Estas pruebas dan valores de probabilidad (ej. p = 0.05) con las que casi todo el mundo ha tenido experiencia en algún momento. Gracias a éstas se pueden hacer aseveraciones como: el grupo “A” es diferente (o no) al grupo “B”; o la temperatura tiene (o no) un efecto en crecimiento (o mortalidad, o color, etc.).

¿Qué son Datos?

Los datos son información que se tiene acerca del mundo. En Biología normalmente los datos hacen referencia a los organismos y su medio ambiente: observaciones que se hacen de características o variables de interés para quien hace el estudio o para la comunidad humana.

Los datos pueden tomar muchas formas y estructuras. Su forma y estructura determina o restringe en gran medida la manera en la cual podrán ser analizados más tarde. Por eso es importante comenzar con una discusión sobre las diferentes clases de datos existentes.

Los datos pueden ser clasificados como **continuos** o **discontinuos** (**discretos**). Los continuos son siempre numéricos y capaces de tomar todos los valores posibles (normalmente dentro de ciertos máximos y mínimos). Como ejemplo consideremos el peso: Dado que el peso no puede ser nunca menor que cero, algo puede pesar 40 kg., 32.46 g., 0.004 mg., etc. En teoría no hay límite para el número de decimales potencialmente requeridos para describir con exactitud el peso de algo. Sin embargo, en la práctica se está siempre limitado por la precisión con la que se puede medir el peso, dado que la tecnología también tiene sus limitaciones.

Por otro lado, los datos discontinuos están restringidos a un grupo predeterminado de posibles valores. Los datos discretos pueden ser o numéricos o alfanuméricos (aquellos valores de una variable expresados con palabras y no con números). Comúnmente se encuentran tales datos en forma de frecuencias, que son números de observaciones y son por ello valores enteros ¡nadie ha visto nunca media observación!). Los datos discretos también pueden ser generados cuando la precisión de la medida de una variable teóricamente continua es muy limitada. A partir del ejemplo anterior con pesos, podemos expresar los datos discretamente como liviano, mediano y pesado, teniendo solamente claves muy crudas de peso. Otros ejemplos de datos discretos son color (rojo, verde, azul, etc.), sexo (masculino y/o femenino) y edad (pichón, juvenil, adulto).

Escalas de datos:

Los datos pueden ser clasificados también de acuerdo al tipo de escala a la que pertenecen. Una escala de datos es un marco de referencia que describe la relación entre valores tomados de la misma escala.

**Escala Radial**: Los datos de este tipo de escala tienen dos características en común. Primero, la diferencia entre cualquier par de unidades adyacentes es la misma, es decir hay un valor de intervalo constante. Por ejemplo, la diferencia entre 90 kg. y 60 kg. es la misma que la diferencia entre 1030 kg. y 1000 kg.. De la misma manera los radios son constantes (ej. 10/20 cm = 50/100 cm). Segundo, todas las escalas radiales contienen un valor de cero real y físicamente significativo. Ej. 0 kg., 0 km/h, 0 s, etc.

**Escala de Intervalos**: Esta escala es igual a la radial excepto en que la selección del cero es arbitraria. Así, éste carece del significado físico del cero en la escala radial. Fíjese que mientras todavía existe un valor de intervalo constante los radios no lo son: 20 °C no es dos veces más caliente que 10°C. Otros ejemplos de escalas de intervalo son las medidas circulares como el tiempo y la dirección de la brújula.

**Escala Ordinal**: Con esta escala ya no se tiene la garantía de intervalos constantes sino que se conoce que sus valores tienen un orden constante. Por ejemplo, en una escala arbitraria los valores A, B, C, D,..., se puede decir que A>B>C>D>... Tales escalas deben ser también internamente consistentes, de modo que si A>B y B>C, debe ocurrir que A>C.

**Escala Nominal**: Estas escalas son usadas en casos en los cuales una característica es clasificada de acuerdo a una cualidad o atributo. Pueden ser números pero solo en el sentido de caracteres **alfanuméricos:** los datos son siempre no numéricos. Diferentes valores en una escala son considerados como tales pero no se puede asegurar que un valor es mayor (menor) o mejor (peor) que otro. Ejemplos de escalas nominales son: color, sexo, vivo, muerto y especies.

Como se mencionó anteriormente, la naturaleza de los datos determinará en gran medida los tipos de análisis que pueden ser realizados. Muchos tipos de análisis son restringidos a ciertos tipos de escalas o pueden ser aplicados sólo a datos continuos o discontinuos. Es muy común tratar de analizar datos solo para encontrarse con que la forma de recolección obstaculiza, inhibe o aun previene el análisis apropiado de los mismos. Esto nos lleva a otro punto importante de este libro:

**¡Pensar por adelantado!**

Estadísticas Descriptivas Numéricas

Las siguientes estadísticas serán aplicables a todos los datos de la escala radial y de intervalos, así como a algunos datos de las escalas ordinal y nominal.

Poblaciones y Muestras:

Estadísticos numéricos son conocidos como parámetros y son propiedades de las poblaciones. Se considera **población** un grupo de organismos o cosas bajo estudio. Una población puede ser muy grande (ej. toda la humanidad, todos los perros de las Américas, todo el volumen de agua del Océano Pacífico) o muy pequeña (ej. todos los ecuatorianos pelirrojos, zurdos, de 210 cm de estatura). El punto importante es que una población es el grupo que se está estudiando acerca del cual se quiere sacar conclusiones. Se discutirá más sobre poblaciones cuando se hable de estadística inferencial.

Hay dos tipos de parámetros. Unos son conocidos **parámetros poblacionales**, los cuales son valores *reales* de la población. En una población dada, puede haber solo un valor verdadero para un parámetro dado. Otros son los conocidos como **parámetros muéstrales** **estadísticos**. Esta segunda clase de parámetros existe porque normalmente somos incapaces de tomar datos sobre todos los miembros de una población. Al no poder hacerlo, es imposible saber los valores reales de los parámetros de esa población. Como consecuencia es necesario estimar estos parámetros de la población en la que se está interesado en base a muestras tomadas al azar.

Generalmente se usan símbolos griegos para los parámetros poblacionales y letras del alfabeto para los parámetros muéstrales correspondientes.

En estadística se asume por conveniencia que el tamaño original de la población (N) es infinito o al menos **muy grande**. En la práctica esto puede no ser cierto y en el caso donde N es muy pequeño (<1000) se requiere el uso de estadísticos especialmente modificados. De igual manera, cuando se habla de seleccionar al azar muestras de una población, normalmente se asume que el tamaño de muestra es relativamente pequeño). Cuando el tamaño de la muestra representa una proporción no muy pequeño de la población es necesario modificar el estadístico que se está usando.

El símbolo 

En estadística hay muchos símbolos que se usa con mucha frecuencia. Uno de ellos es el símbolo Σ. Este símbolo significa sumar. Por ejemplo, para indicar la sumatoria de una serie de números (X1+X2+...+Xn), se escribe:

 (1.1)

Se entiende  como la sumatoria de los números Xi desde X1 a Xn

A menudo, se escribe  como  y significa lo mismo.

Medidas de Tendencia Central:

Son los estadísticos más importantes y comunes que se encuentran. Una medida de Tendencia Central (**TC**, téngase presente esta abreviación pues se la verá frecuentemente a lo largo de este libro) es una descripción de como es el individuo promedio o típico de una población. La medida de TC más común es la **media aritmética**. El símbolo para la media aritmética de la población es **µ** y el de una muestra es. Recuerde que la media no es la única medida de TC y no es siempre la mejor o la más apropiada.

La media aritmética de una población (µ) se define como:

 (1.2)

Donde, es la suma de todas las observaciones (X1+X2+...+XN);

N es el tamaño de la población.

La media aritmética de una muestra () se define como:

 (1.3)

Donde, es la suma de todas las observaciones (X1+X2+...+Xn);

n es el tamaño de la población.

Es importante que se entiendan las fórmulas precedentes, porque serán utilizados varias veces a lo largo de este libro.

Se dice que  es el estimativo más **eficiente**, **no sesgado** y **consistente** de la µ. Por eficiente entendemos que la  de la muestra usa la mayor cantidad de información posible de los datos para estimar µ. Se dice que no es sesgado porque no tiende a sobre o subestimar µ, no importa la naturaleza de la población original. Finalmente, se llama  consistente porque generalmente, aunque no siempre, da un buen valor estimativo de µ.

Otro parámetro de TC es la tendencia **mediana.** La mediana de la población tiene el símbolo **M**; la mediana de una muestra no tiene un símbolo aceptado. La mediana de una muestra es el mejor estimativo de la mediana de la población y es también un estimativo consistente y no sesgado de µ (sin embargo no es tan eficiente como). La mediana es el valor o medida del medio en nuestros datos. Así, en cualquier grupo de datos hay tantos valores mayores a la mediana como hay menores que ella.

La mediana de la muestra es calculada ordenando primero los datos de acuerdo al tamaño (ascendente o descendente) y calculando entonces la mediana como:

 (cuando n es impar) (1.4a)

 (cuando n es par) (1.4b)

Donde X es el valor para cada observación.

En palabras: Cuando hay un número impar de observaciones ordenadas (número de observaciones = n) la mediana es igual a la observación (n+1)/2. Cuando n es par, entonces la mediana es igual al promedio de la observación número n/2 más la observación número (n/2)+1.

Cuando comparamos la mediana y la media observamos lo siguiente:

1. La mediana usa menos información a partir de los datos que la media, ya que usa solamente el orden de los datos y no estos mismos. Así, la mediana es menos eficiente.
2. La mediana es menos afectada por valores extremos que la media. Se dice por ello que la mediana es un estadístico más **robusto** que la media. Esta es a veces una característica muy deseada.
3. La mediana es a menudo más apropiada cuando hay gran error de medición asociado con los datos.
4. La mediana puede ser usada con datos ordinales mientras que la media resulta inapropiada en tal caso.

Otras Medidas de Tendencia Central:

La mediana y la media son las medidas de TC de uso más frecuente. Otras medidas son usadas con menor frecuencia en circunstancias donde son apropiadas.

- La **Moda** es descrita como el valor que ocurre con más frecuencia. En ocasiones puede haber más de una moda. Esta medida es a menudo subjetiva en alguna forma. La moda de la muestra es el mejor estimativo de la moda de la población. Es muy robusta contra los valores extremos, pero muy sensible al tamaño de la población y es un estimador sesgado e ineficiente de µ. Por lo general no se usa la moda en Biología.

- El **Punto medio** **o Rango medio** (dos nombres para denotar lo mismo) es el punto medio entre los valores máximo y mínimo. Se define como (mínimo + máximo)/2. Esta medida es un estimador muy pobre e ineficiente de µ. Comúnmente se encuentra asociado a datos de temperatura.

- La **Media Geométrica** (**Gm**) se usa a veces cuando se promedian radios y porcentajes. Se define como:

 (Población) (1.5a)

 (Muestra) (1.5b)

El símbolo Π se define como el producto de todas las observaciones.

- La **Media Armónica** (**Hm**) es usada algunas veces cuando se promedian tasas. Se defina como:

 (Población) (1.6a)

 (Muestra) (1.6b)

Nótese que Hm no tiene interpretación donde hay ceros en los datos (¿Por qué?). Ambas, la Gm y la Hm pueden ser aplicadas a datos de escalas radiales.

Medidas de Dispersión y Variabilidad:

Muy a menudo es de interés conocer o describir, además de la TC, la dispersión o variación alrededor de un valor. Las medidas de dispersión describen parámetros poblacionales y nosotros los estimamos con valores muéstrales.

El **Rango** es la distancia entre los valores máximo y mínimo. Provee de muy poca información acerca de la dispersión. Más aún, los estimativos del rango tienden a subestimar el valor del verdadero rango. Así, el rango de la muestra es un estimador sesgado del rango de la población.

Un parámetro mucho mejor para describir la variación es **Desviación Promedio**. Esta puede ser interpretada como la diferencia promedio positiva entre alguna medida de TC y cada valor en la muestra (o población). La desviación promedio se define como:

 (1.7)

Donde |Xi-TC| se lee “valor absoluto de la diferencia entre cada observación y una medida de TC” (usualmente media o mediana).

El uso del valor absoluto es necesario ya que por definición la suma de las desviaciones (donde TC es la media) será cero.

Al llegar a este punto se presentará una nueva medida de desviación llamada la **Suma de los Cuadrados** (SC) de la población, la cual se define como:

 (1.8)

De la misma forma, la suma de los cuadrados de una muestra se define como:

 (1.9)

Obsérvese la similitud entre la suma de desviaciones y la suma de cuadrados. En ambos casos se está calculando una suma positiva de las diferencias entre la media y cada observación. La analogía puede tomarse un paso más adelante cuando se calcula la media de las desviaciones al cuadrado. Esta cantidad es conocida como **Varianza**. Matemáticamente la varianza es:

 (Población) (1.10a)

 (Muestra) (1.10b)

La varianza siempre tiene unidades al cuadrado. Por ejemplo, si se mide en km. las unidades de la varianza serán km2. Se usa **n-1** para la varianza de la muestra porque ésta tiende a subestimar la varianza con **n**. Así, podría ser un estimador sesgado de la varianza de la población. Por otro lado, si dividimos entre **n-1** la varianza no es sesgada. En este contexto se conoce a n-1 como los **grados de libertad**. Se tratará esto de nuevo más adelante.

La fórmula para SC dada arriba es ineficiente porque se necesita hacer dos recorridos a los datos. Uno para calcular la  y otro para calcular cada una de las desviaciones al cuadrado. Una fórmula mucho más eficiente (conocida como la fórmula de calculadora, ya que es usada para computadoras) es:

 (1.11)

La Medidas de Dispersión se puede expresar entonces de la siguiente manera:

 (1.12)

Para tener una medida de variación con las mismas unidades de los datos originales, se saca la raíz cuadrada de la varianza. Esta medida es llamada la **Desviación Estándar**:

 (Población) (1.13a)

 (Muestra) (1.13b)

Algunas veces es deseable hacer comparaciones de la variación entre diferentes poblaciones. Como las magnitudes de s2 y s (Σσ2 y σΣ) dependen de la magnitud de los propios datos, si se dividen entre  se da origen a un valor adimensional, corregido para el efecto de magnitud. Este valor es **Coeficiente de Variación (V)**.

 (1.14a)

 (1.14b)

Para demostrar las estadísticas que acabamos de discutir usaremos los datos obtenidos de un estudio pequeño de 25 personas. Cada persona fue medida por su altura y los datos se encuentran abajo.

**Alturas (m)**

(1) 1.47 (2) 1.47 (3) 1.48 (4) 1.50 (5) 1.53

(6) 1.55 (7) 1.56 (8) 1.56 (9) 1.60 (10) 1.61

(11) 1.63 (12) 1.64 (13) 1.65 (14) 1.65 (15) 1.65

(16) 1.66 (17) 1.66 (18) 1.67 (19) 1.67 (20) 1.69

(21) 1.70 (22) 1.73 (23) 1.74 (24) 1.78 (25) 1.81

n = 25 n es impar





















Capítulo II

Introducción a los Gráficos (1)

Introducción

¿Qué son Gráficos?

En este libro se considera cualquier representación visual de los datos como un gráfico. Como se verá, esta definición incluye un rango muy amplio de técnicas y estilos. Al igual que en estadística descriptiva numérica, el aspecto más difícil en gráficas no es aprender las diferentes técnicas sino saber cuál será la más adecuada para exponer el problema en el que se está trabajando.

¿Por qué se usan Gráficos?

Las tablas de datos son difíciles de interpretar cuando tienen más de unos pocos números. Por otro lado, no hay límite para la cantidad de datos que pueden ser procesados gráficamente: fotografía común está compuesta en realidad, por millones de piezas individuales de la información. ¡¡La marca actual para la mayor cantidad de información (datos) presentados en un gráfico interpretable es de aproximadamente un millón de números!!

Se verá que mientras los gráficos pueden ser herramientas descriptivas muy poderosas, será incomprensible si se intenta representar un tipo errado y/o demasiados datos. El tipo y la cantidad de datos que pueden ser representados en un gráfico particular dependerán del tipo de gráfico usado, la naturaleza de los datos y la habilidad e imaginación de la persona que lo crea.

Los gráficos sirven también como herramienta analítica. Con su ayuda se pueden detectar patrones y relaciones entre los datos que podrían ser demasiadas sutiles para ser detectados por pruebas estadísticas, o para los cuales no existen tales pruebas. Esto es especialmente cierto cuando se está tratando con datos que tienen componentes espaciales o temporales determinantes.

En resumen, los gráficos son una parte muy importante y poderosa de la estadística, tanto por sus capacidades descriptivas como analíticas. Como regla general es recomendable comenzar cualquier análisis de datos mirándolos en su representación gráfica. En el peor de los casos un gráfico dará una mejor idea acerca de los datos y en el mejor mostrará aspectos que nunca se sospechó que existieran.

Las técnicas que se discutirán pueden ser clasificadas en función de los tipos de datos y preguntas que se pueden hacer sobre ellos. Normalmente estos son los usos más comunes para cada una de las técnicas, pero siempre hay muchos usos atípicos para cada técnica.

Técnicas para Graficar Una o Dos Variables

Variables discretas basadas en escalas nominales u ordinales:

Pregunta: ¿Cuántas? ¿En qué proporción?

Los datos que caen en estas categorías son típicamente conteos por observaciones de individuos que pertenecen a una o varias clases discretas. Por ejemplo: el número de hembras y machos, el número de individuos observados comportándose de cierta forma. Muchas veces se crean clases discretas a partir de variables teóricamente continuas. Con más frecuencia esto ocurre debido a un alto grado de imprecisión en las medidas (ej. tamaño: pequeño, mediano y grande).

Como ejemplo se considera un caso que incluye la captura de iguanas vivas. La edad aproximada de cada una es estimada y luego liberada. Dado que es muy difícil estimar la edad absoluta de las iguanas, la mejor forma de hacerlo es clasificarlas de acuerdo a una de cuatro clases muy generales. Los resultados de este experimento son como sigue:

Tabla 2.1

Grupo de Edad Número de capturados % del Total

Recién Eclosionadas 63 35

Juveniles 36 20

Sub-adultas 27 15

Adultas 54 30

180 100

Para responder a las preguntas “¿Cuántos?” y “¿En qué proporción?” el gráfico más usado es el gráfico **diagrama de barras**. En éste la frecuencia o número de observaciones asignado a cada clase (en nuestro ejemplo el número de iguanas asignado a cada grupo de edad) es representado por la altura de una barra diferente. Para los datos anteriores podemos construir el diagrama de barras siguiente:

**Gráfico 2.1**



Este es el estilo más simple y común de diagrama de barras. Las modificaciones posibles incluyen una orientación horizontal, escala de ejes no lineales y barras tridimensionales.

Supóngase que ahora se desea comparar estos resultados con otros de estudios anteriores. Se desea comparar la proporción en cada grupo de edad en vez de los números absolutos ya que el número total de iguanas capturadas cada vez fue diferente. Por lo tanto es necesario convertir las frecuencias absolutas en frecuencias relativas (%). Esto se logra calculando el porcentaje del total que cada una representa. Para la iésima clase definimos el porcentaje de frecuencia como:

 (2.1)

Donde: Xi = Número de observaciones en clase i

n = Número total de observaciones en todas las clases.

Con los mismos datos usados en la Tabla 2.1 se obtendría:

**Gráfico 2.2**



Nótese que la forma relativa de las barras no cambia y que la altura absoluta depende solamente del máximo valor escogido para el eje vertical.

Otra forma de representar los mismos datos de frecuencia es el **Diagrama de Pastel** (**Pie**). Este es más apropiado cuando se desea representar datos que agrupados corresponden a un número total de observaciones o individuos. Estos diagramas son muy útiles para mostrar la contribución relativa hecha por cada clase, pero no dan información acerca de los números absolutos, a menos que se aumente con descripciones numéricas.

Un diagrama de pastel es sencillamente un círculo dividido en secciones que representan el porcentaje de frecuencias de cada clase. El área de cualquier sección es directamente proporcional al porcentaje de frecuencia de otra clase. Para los mismos datos de captura de iguanas el diagrama de pastel sería:

**Gráfico 2.3**

****

Los diagramas de pastel son efectivos cuando se tiene un número de clases relativamente pequeño. Cuando hay más de diez clases éste tiende a ser difícil de interpretar y comparar. Frecuentemente, como se ve en el Gráfico 2.3, una o dos secciones del diagrama sobresalen del resto para enfatizarlas.

Variables continuas basadas en escalas de intervalos o radiales:

Pregunta: ¿Cuántas? ¿Cómo están distribuidas?

En Biología con frecuencia se quiere conocer cómo se distribuyen los datos basados en variables continuas. Para hacer esto a menudo se dividen las variables continuas (peso, altura, edad) en intervalos discretos de la misma amplitud.

Para representar gráficamente estos datos se usa el **Histograma**. Los pasos involucrados en la creación de un histograma son los que siguen:

1. Encontrar los valores máximo y mínimo del grupo de datos.

2. Escoger los valores máximo y mínimo más convenientes para el histograma.

3. Escoger la amplitud de clase a usarse y calcular el número de intervalos, ó

Escoger el número de intervalos y luego calcular la amplitud de cada clase.

4. Calcular el rango para cada intervalo.

5. Calcular la frecuencia para cada intervalo.

6. Dibujar el gráfico.

7. Examinar el gráfico, y en caso de encontrar muchos intervalos vacíos en el, volver a revisar pasos 2 y/o 3.

En el punto 3 se debe considerar que es necesario contar en el número de clases suficientes para ser informativo, pero no con tantas que muchos intervalos no tengan observaciones, especialmente si estos están localizados en la mitad del gráfico.

Como ejemplo, volvamos a nuestras iguanas capturadas. Para cada una de las 180 iguanas se tiene medida de su peso. Interesa ver como se distribuyen los pesos.

1. El peso máximo observado es de 5.4 kg. y la más pequeña de 0.2 kg.

2. Se escoge el valor mínimo para el gráfico como 0 kg. y el valor máximo de 5.5 kg.

3, 4, 5. El tamaño de intervalo de 0.5 kg. Entonces habrán intervalos de la siguiente manera: 0 a < 0.5, 0.5 a < 1.0, 1.0 a < 1.5,..., 5.0 a < 5.5. El símbolo < debe interpretarse como: “Todos los valores hasta este, sin incluir el número siguiente.”

6. El histograma resultante es el que se muestra en el Gráfico 2.4):

**Gráfico 2.4**

****

Un problema que siempre ocurre con histogramas es decidir a qué clase pertenece una observación dada. Esto sólo se convierte en problema cuando un valor corresponde exactamente al límite entre dos intervalos. En el caso del 2.4, ¿En qué intervalo caería una observación de 3.0 kg.? La respuesta a esta pregunta depende de cómo se definen los intervalos antes de construir el histograma. Se puede declarar específicamente que cualquier intervalo incluye todos los valores comenzando en un cierto valor y extendiéndose hasta otro, pero sin incluirlo, como en el ejemplo anterior; o se puede definir el intervalo para que incluya todas las observaciones mayores que algún valor hasta otro menor incluyéndolo. Generalmente no importa cual de los métodos se usa, si se establece claramente y se emplea consistentemente.

El mayor uso de los histogramas con datos continuos es revelar como se distribuyen las observaciones. Estos son especialmente útiles para indicar el grado de variabilidad y la simetría en los datos. En Biología se está especialmente interesado en ver si los datos corresponden a alguna distribución teórica tal como la binomial, normal o Poisson. Al graficar los datos junto a una distribución teórica se pueden determinar rápidamente si los datos son marcadamente diferentes de lo esperado. Por ejemplo, muchas veces los fenómenos biológicos se expresan en datos distribuidos normalmente. Puede ser que se desee conocer si el peso de las iguanas capturadas tiene tal distribución. El Gráfico 2.5 es una comparación de los datos del ejemplo con los de una distribución normal (línea punteada) con la misma media y varianza.

Se puede ver en el Gráfico 2.5 que hay una diferencia considerable entre la distribución de nuestros datos y una distribución normal. Lógicamente el próximo paso sería hacer una prueba para ver si esta diferencia cae en el rango que normalmente se esperaría debido a la variación causada al azar en el muestreo. Desafortunadamente estas pruebas están fuera del alcance de este libro. Para mayor información referirse a los libros que presentamos en la bibliografía.

**Gráfico 2.5**



Un segundo y más efectivo método para comparar distribuciones es el llamado **Histograma de Frecuencias Acumulativas** o **Histograma FA**. Este es muy similar al discutido anteriormente, con la única modificación de que cualquier intervalo es la suma de todas las observaciones que pertenecen a tal intervalo y a todos aquellos menores (o a la izquierda). Es decir, es la suma acumulativa de todos los intervalos previos. Matemáticamente el iésimo intervalo de un histograma FA se define como:

 (2.2)

Donde F(j) es el número de observaciones del intervalo j.

La mayoría de histogramas FA se construyen usando **Frecuencias Acumulativas Relativas** (**FAR**). En el caso FAR, el porcentaje de todas las observaciones es representado por la frecuencia acumulativa de un intervalo dado. Para el intervalo iésimo, la FAR se define como:

 (2.3)

Donde n es el número total de observaciones.

Para construir un histograma FAR se necesita calcular primero la FAR para cada intervalo. Estos valores se muestran en la Tabla 2.2.

**Tabla 2.2**

I Rango F(i) FA(i) FR(i) FAR(i)

1 0.0 - 0.5 2 2 0.011 0.011

2 0.5 - 1.0 5 7 0.028 0.039

3 1.0 - 1.5 10 17 0.056 0.094

4 1.5 - 2.0 18 35 0.100 0.194

5 2.0 - 2.5 30 65 0.167 0.361

6 2.5 - 3.0 25 90 0.139 0.500

7 3.0 - 3.5 28 118 0.156 0.656

8 3.5 - 4.0 23 141 0.128 0.783

9 4.0 - 4.5 20 161 0.111 0.894

10 4.5 - 5.0 14 175 0.078 0.972

11 5.0 - 5.5 5 180 0.028 1.000

La Tabla 2.2 se representa en el Gráfico 2.6. La línea punteada corresponde a la curva normal.

**Gráfico 2.6**



Las mayores desviaciones de las distribuciones propuestas son reconocidas con mayor facilidad usando histogramas FAR que con histogramas normales. Con un poco de práctica es muy fácil reconocer diferencias significativas.

Cualquier variable continúa dividida en sub-grupos:

Pregunta: ¿Cuáles son las diferencias entre los sub-grupos?

Muy a menudo se tienen datos (tales como el peso de las iguanas) que constituyen variables continuas pero que pueden subdividirse en categorías discretas (como sexo, localización de captura, especie, etc.). En este caso, a menudo interesa hacer una comparación entre categorías de ciertos parámetros claves (tales como media y error típico o error estándar) entre los sub-grupos. Si el número de sub-grupos es pequeño, las comparaciones se pueden hacer mejor mediante la elaboración de una Tabla simple, como la Tabla 2.3.

**Tabla 2.3**

Isla peso Desv.Est. Mínimo Máximo

Santa Cruz 3.4 1.5 1.5 5.2

Baltra 3.1 1.1 1.4 4.5

Seymour Norte 3.5 1.4 1.8 5.3

Española 4.0 1.8 1.3 6.1

Daphne 3.0 0.9 2.0 4.0

Isabela 3.8 1.5 1.9 5.2

Fernandina 2.5 1.4 1.0 4.7

Estos datos se pueden representar parcialmente en un gráfico adaptando una técnica usada anteriormente: el Diagrama de Barras. En vez de representar la frecuencia con la altura de una barra, ésta puede representar un valor absoluto, que en nuestro caso puede ser la media de los pesos. Cada barra representará la media del peso para una isla. También se puede indicar la desviación estándar o el rango de los datos añadiendo una barra delgada en la parte superior de cada barra. En el caso de la desviación estándar la barra delgada se añade a la parte de arriba para indicar la media más una desviación estándar (Gráfico. 2.7a). El rango de los datos será representado por una línea delgada más larga, con la media localizada en el lugar que ocupa en relación al rango (Gráfico. 2.7b).

Aunque el Gráfico 2.7 permite comparar las medias de los sub-grupos rápidamente, está limitada en que máximo muestra una medida de TC (véase el Capítulo I) y una de dispersión. Una técnica para graficar usada comúnmente para los mismos datos es llamada **Diagrama de Caja y Barras** o **Dice**. Estos permiten usar una medida de TC y hasta tres de dispersión.

**Gráfico 2.7**



Hay muchos estilos de diagramas de caja y barras dependiendo de qué combinación de parámetros van a ser representados. Todos los diagramas de caja siguen el mismo patrón general mostrado en el Gráfico 2.8.

**Gráfico 2.8**



Normalmente las cajas se dibujan verticalmente y están orientadas con respecto a una escala localizada en el eje vertical. Para dibujar una caja, se busca Tendencia Central en el eje vertical y se representa por una barra horizontal delgada (Gráfico 2.9a). Después se dibuja la caja alrededor de la Tendencia Central (media, mediana, etc.) ± una unidad de dispersión (desviación estándar, varianza, cuartiles de 25/75, etc.). Véase el Gráfico 2.9b. Finalmente se añade una barra vertical delgada para indicar otra clase de dispersión (el rango, máximo y mínimo). Véase el Gráfico 2.9c.

**Gráfico 2.9**



Hay muchas combinaciones posibles de datos que pueden ser representados en diagramas de caja. Las más comunes son:

- Media, Desviación Estándar, Mínimo y Máximo.

- Media, Desviación Estándar, Intervalos de Confianza.

- Mediana, Cuartiles 25-75%, Mínimo y Máximo.

La combinación de datos que se usen dependerá de la preferencia personal y del tipo de pruebas estadísticas que va a aplicar a sus datos. Esta técnica es muy útil para resumir datos que van a ser analizados después usando un **Análisis de Varianza** (**ANOVA**). Los datos de la Tabla 2.3 se muestran en el Gráfico 2.10.

**Gráfico 2.10**

****

Datos pertenecientes a escalas radiales o de intervalos.

Pregunta: ¿Cuál es la relación entre dos variables?

Muchas veces en Biología se tienen datos que consisten en pares de observaciones relacionadas al mismo individuo, momento, área, etc., y se desea conocer si hay algún tipo de relación entre las observaciones. En otras palabras, si hay cualquier clase de correlación o relación causal entre las dos observaciones. Muy a menudo las observaciones representan diferentes variables y es importante conocer si las dos variables están relacionadas de alguna manera. Por ejemplo, un par de observaciones pueden consistir en medidas de peso y altura de un individuo. La pregunta de interés podría ser: ¿Están el peso y la altura de un individuo relacionados? y si es así, ¿Cómo se pueden describir matemáticamente?

Ya es obvio que la técnica a usar es el **gráfico bidimensional simple (2D)**. El proceso de crear gráficos 2D simples, comienza con la decisión acerca de cuál variable debe ser representada por el eje vertical y cuál por el horizontal. En muchos casos esto no importa: por ejemplo, casos en los que no hay razón para sospechar que una variable está controlando o influyendo a la otra y sólo deseamos ver si hay una correlación entre las magnitudes de las dos. En el caso de haber una razón *a priori* para sospechar de una relación causal, o donde deseamos ser capaces de describir la magnitud de una variable en términos de una función de la otra, se coloca tradicionalmente la **variable independiente** en el eje horizontal y la **variable dependiente** en el vertical. La variable dependiente es aquella que es dependiente de (o causada por) la otra variable (independiente). Por ejemplo, el peso (variable dependiente) de un individuo puede ser una función de (o causada por) la edad (variable independiente).

El paso siguiente en la construcción de un gráfico 2D es encontrar los valores máximo y mínimo para ambas variables. A partir de éstos se escogen los valores máximo y mínimo para los ejes. Luego se escoge el número de intervalos y el tamaño que se va a usar en los ejes (muchos o muy pocos intervalos, hacen difícil de leer un gráfico). Finalmente, se necesita localizar cada par de observaciones en el gráfico.

Como ejemplo, los datos de la Tabla 2.4 se usarán para crear un gráfico 2D simple (Gráfico. 2.11).

**Tabla 2.4**

No. de Individuo Altura (cm.) Peso (g)

1 160 55

2 180 70

3 170 60

4 165 62

5 190 88

6 175 67

7 174 60

8 183 78

9 192 100

10 195 110

máx. 195 110

mín. 160 55

El Gráfico 2.11 es posiblemente el tipo más sencillo de gráfico de 2D. Hay muchas modificaciones posibles de este gráfico. Por ejemplo, cuando una variable tiene un componente secuencial, como las observaciones hechas a lo largo del tiempo, es común conectar los puntos en el gráfico (Gráfico 2.11).

**Gráfico 2.11**



**Gráfico 2.12**



A veces**,** cuando se tienen muchas observaciones hechas para un número relativamente pequeño de condiciones (Gráfico 2.13a), se resumen los datos indicando solamente la media y alguna medida de dispersión (Gráfico 2.13b).

En el Gráfico 2.13a nos gustaría explorar la relación entre el peso de huevos a diferentes temperaturas. Debido a que los datos están tan dispersos es difícil discernir las diferencias en las temperaturas. Cuando se resumen los datos como en el Gráfico 2.13b, las diferencias en el peso de huevos debido a las condiciones de temperatura son más notorias. Fíjese que el Gráfico 2.13b pudo haber sido representado en un diagrama de caja. Dado que la temperatura pertenece a una escala de intervalos más que a una nominal, se ha usado un gráfico 2D. Al final, el tipo de pruebas estadísticas que se usarán determinará el método más apropiado.

**Gráfico 2.13**



Otro aspecto de estos gráficos que es cambiado a menudo, es la escala de los ejes. En las Gráficos 2.11, 2.12 y 2.13, todos los ejes tienen un ancho de intervalo constante. En otras palabras, son lineales. Si el grupo de datos presenta una relación lineal entre dos variables **X** e **Y** cuando se grafican en dos ejes ortogonales (lineales) esta relación tiene la forma:

 (2.4)

En muchas ocasiones dos variables no están relacionadas por una ecuación lineal simple, y por lo tanto no muestran una relación en forma de línea recta cuando se grafican en ejes ortogonales. Algunos ejemplos son:

 (2.5a)

 (2.5b)

 (2.5c)

La graficación de datos con estas relaciones funcionales muestra curvas no lineales y a veces son muy difíciles de graficar en ejes ortogonales. Por ello, se recurre a **transformaciones**. Estas son técnicas que permiten mostrar relaciones lineales entre datos que están relacionados en formas no lineales a través del cambio de escala en los ejes. Los ejes no lineales más comunes son los **logarítmicos** y a la transformación que hacemos se le llama **logarítmica**. Si se sabe de antemano qué forma tiene la relación de las variables se puede establecer qué tipo de transformación hacer para llegar a una línea. Por ejemplo, relaciones de la forma  **y = αxβ** pueden transformarse en lineales después de hacer una transformación **log(x), log(y)**; una de la forma **y = αeßx** o **y = αcßx** con una **log(x)** (**c** es una constante). Los logaritmos naturales y comunes pueden ser usados indistintamente con los mismos resultados. Como un ejemplo las Gráficos 2.14a y 2.14b, grafican las variables **X, Y** de la Tabla 2.5 en ejes ortogonales y logarítmicos respectivamente.

**Tabla 2.5**

X Y Log(X) Log(Y)

1 0.1 0.00 -1.00

2 0.8 0.30 -0.10

3 2.7 0.48 0.43

5 12.5 0.70 1.10

10 100.0 1.00 2.00

15 337.5 1.18 2.53

20 800.0 1.30 2.90

**Gráfico 2.14**



Una vez que se han graficado los datos en ejes logarítmicos se pueden calcular los valores de **α** y **ß** (véase los Capítulos 6 y 7).Si se desconoce la función que relaciona las variables con las que se está trabajando, puede ser necesario graficar los datos en los dos tipos de ejes, log-log y log-lineal.

Un segundo método de representar las variables **X** e **Y** de la Tabla 2.5 es graficar los valores logarítmicos transformados en ejes lineales. Esto tiene el mismo efecto que graficar los valores originales en ejes logarítmicos (véase el Gráfico 2.14c).

No se pueden hacer transformaciones logarítmicas cuando los datos son menores o iguales a cero (¿por qué?). Si sus datos contienen puros valores negativos se deben graficar los valores absolutos. En caso que los datos contengan ceros, los datos se pueden transformar usando **log(x+1)** o **log(y+1)** dependiendo del eje dónde se encuentran los valores iguales o menores a cero. Si se usa esta relación y se desea calcular **α** y **ß** se debe corregir por el **+1** (esto se verá en el Capítulo VII).