# TALLER DE ESTADISTICA APLICADA A LA ACUICULTURA

Fabrizio Marcillo Morla
Cámara Nacional De Acuicultura

**Introducción Teórica.-** (1 hora)

Introducción

Buenos días, mi nombre es Fabrizio Marcillo, y seré su instructor a lo largo de este curso.
En primer lugar les explicaré el objetivo de este curso, y en que forma sería bueno que lo entiendan.

El curso es una revisión de varios procesos y métodos estadísticos expresados de una manera práctica e inteligible.

No es intención de este curso dar una explicación detallada y profunda de los procesos matemáticos que involucran las suposiciones y cálculos estadísticos, sino más bien dar una guía práctica sobre los cálculos a realizar en determinada situación, dando una idea general de lo que es diseño experimental.

Los métodos descritos en este curso son de estadística general, y por lo tanto pueden ser usados en distintas situaciones; sin embargo, como el curso está programado especialmente para ser usado en acuicultura, los ejemplos que se darán corresponden a esta área.

Los tópicos a tratar entrarán dentro de 4 grupos principales, esto es:

- Conceptos y definiciones básicas de estadística, así como ciertas suposiciones que vamos a hacer.

- Teoría de probabilidades y distribuciones teóricas de las mismas.

- Diseño experimental, enfocado con lógica y sentido común, y basados en los conceptos básicos y teoría que hallamos estudiado.

- Cálculos aritméticos y tablas numéricas, que realizados de forma rutinaria y mecánica después de haber definido nuestro problema, nos darán los materiales sobre los cuales se van a basar las inferencias y medir la incertidumbre.

El objetivo del curso es dar a Uds. conocimientos básicos de Estadística aplicada, esto es saber reconocer el problema que tienen adelante y aplicar los cálculos aritméticos correspondientes para obtener una respuesta. Y, fomentar una forma de pensar clara y disciplinada, especialmente cuando se trata de recolectar e interpretar información numérica.

Con las bases y la comprensión obtenidas en este curso se espera que Uds. puedan ampliar estos procesos de acuerdo a sus necesidades individuales consultando obras más avanzadas.
Deben de tener en cuenta que los procesos que aquí se describen son herramientas de trabajo, y podrán como ellas ser utilizadas de acuerdo a su juicio y necesidades. Aquí se darán ejemplos de algunas maneras que se pueden utilizar estas herramientas, pero no son las únicas aplicaciones de estas, y , ni siquiera las mas representativas en muchos casos.

Como herramientas que son, la mayoría de los procesos que se describen pueden ser utilizadas para ayudar a tomar decisiones, pero todo proceso de toma de decisiones necesita de un análisis racional para poder determinar la conveniencia de aplicar determinada alternativa. Las soluciones que se obtengan con estas herramientas son pautas que nos ayudarán a seleccionar dicha alternativa.

Al final de cada sección se hará una revisión o estudio de casos con el objetivo de poner en práctica los puntos revisados. Estos casos son situaciones prácticas (no todos reales) que <u>no necesariamente deben de tener una sola respuesta</u>, y que no deben de ser tomados como ejemplos de buena práctica en Acuicultura, si no mas bien como problemas a resolver desde el punto de vista didáctico en bioestadística.

En cursos anteriores que dicté, traté de dar mayor énfasis en como realizar los cálculos aritméticos, esto pienso que no es lo más útil, por esa razón nos centraremos más en dar las bases teóricas, probabilísticas y prácticas para poder identificar el problema que tenemos, y que solución aplicar. Se dará algunos ejemplos de como resolver los cálculos aritméticos, y en manual adjunto hay fórmulas para la resolución de una variedad de pruebas, pero se espera que en la práctica la resolución se la haga mediante el uso de algún software. En este curso veremos algunas de las herramientas estadísticas disponibles en Excell, las cuales presentan tablas de resultados bastante completas, y se explicará como interpretar los resultados que salgan en dichas tablas.

El programa es el siguiente:

Empezaremos viendo una serie de definiciones y conceptos básicos, así como los usos de la estadística, veremos las diferencias entre población y muestras y entre estadísticos y parámetros.
Continuaremos con probabilidades y distribuciones, la parte más importante de entender, ya que es la esencia misma de la estadística, les pediré un poco de paciencia con esta parte, ya que tal vez les pueda parecer un poco aburrida, pero si dominan esta parte todo el resto viene por simple lógica.
Después veremos estadística descriptiva, veremos diversas formas de presentar los resultados, distribución de frecuencias, Muestreos y la estimación de parámetros y errores para cada uno de ellos, tamaño de la Muestra, como expresar relaciones entre dos o mas variables con regresión lineal, y como describir dis-

tribuciones de probabilidad de variables muestreales ajustándolas a alguna función de probabilidad conocida.
Veremos luego estadística comparativa, que es la parte que pienso que a la mayoría le interesará. Haremos un breve repaso de la base probabilística de la misma, y entraremos de lleno en las pruebas de hipótesis, como usarlas y cuando usarlas, así como la forma de interpretar los resultados. veremos análisis de varianza (ANOVA) para varios diseños y comparaciones múltiples para medias. Todo esto lo enlazaremos con bases de diseño experimental.
Finalmente revisaremos algo de estadística predictiva, el cual es una parte interesante de a estadística, veremos como ajustar curvas de datos muestreales para poder predecir valores futuros, métodos de series de tiempo para separar diferentes fuentes de variación en un modelo con influencias estacionales y/o cíclicas, cálculo de probabilidades y valor esperado y algunas bases de usos de métodos de simulación.

No entraremos a explicar los procesos matemáticos que justifican la teoría que estudiaremos, dedicándonos exclusivamente a cálculos y teoría con aplicación directa. Si desean el desarrollo matemático de la teoría y fórmulas Uds. pueden encontrarlo en cualquier libro de estadística.

Usos de la Estadística, Definiciones y Conceptos Básicos

Para empezar Definiremos Estadística como: La ciencia pura y aplicada (no exacta), que crea, desarrolla y aplica técnicas de modo que pueda evaluarse la incertidumbre.

Revisemos un poco este concepto:
Ciencia es "un conjunto de conocimientos comprobados y sistematizados".
Pura.- Por cuanto estudia ciertos procesos teóricos.
aplicada.- En cuanto se encarga de resolver problemas específicos.
(no exacta).- Por cuanto con ella no podemos obtener un resultado único, si no probabilidades de resultados esperados.

La misma no es nueva, ya que desde tiempos antiguos se la usaba principalmente en los conocidos censos.
Posteriormente (siglo XVII) se desarrolló la teoría de probabilidades basada en los juegos de azar.
Muchas teorías, principalmente biológicas como las de Mendel o las de Darwin, tuvieron bases estadísticas.
Sin embargo los métodos de la estadística moderna, los cuales se utilizan actualmente, no se desarrollaron hasta mediados del siglo XIX y principios del XX, principalmente para uso en biología, agricultura y genética.
Algo que he oído con frecuencia en las camaroneras es que los principios estadísticos son "matemáticos", y que por lo tanto no se aplican a la biología, pero como Uds. verán, la mayoría de estos métodos fueron desarrollados precisamente para ser usados en biología, y calculados a partir de poblaciones natura-

les y sus muestras, a fin de poder evaluar el "todo" y sus variaciones naturales en base a una parte, con cierto grado de certeza.

El objetivo fundamental de la estadistica es hacer inferencias acerca de una pobación con base en la información contenida en una muestra

**Variables y Estadísticos.-**

<u>Definición de Variables</u>
Llamamos variable a una propiedad con respecto a la cual los individuos de una muestra o una población se diferencian en algo verificable.
En otras palabras son ciertas "características" que presentan variación.

Variables mensurables son aquellas cuyos diferentes valores pueden ser expresados de manera numérica, por ejemplo el peso y la longitud.
Variables ordinales son aquellas que pueden ser expresadas en orden de magnitud; por ejemplo, los grados de madurez o de llenura o los índices de lípidos.

Atributos o variables cualitativas son aquellas variables que no pueden ser medidas, pero pueden ser expresadas cualitativamente. Ellos representan propiedades; por ejemplo, el color, sexo, etc.

Variables discretas son aquellas cuyo conjunto de posibles valores son fijos, y no pueden tomar valores intermedios, por ejemplo el número de peces en un acuario.

Definimos como variables continuas a aquellas cuyo conjunto de posibles valores puede alcanzar un número infinito entre dos valores cuales quiera, por ejemplo la longitud o el peso.

Llamamos variables independientes a aquellas cuyo valor no depende de otra variable.

Llamamos variables dependientes a aquellas cuyo valor va a depender de otra función.

<u>Definición de Datos.</u>
Llamamos datos u observaciones a cualquier valor numérico o cualitativo que mida una variable. En otras palabras, a los valores experimentales que va a tomar una variable determinada.

<u>Población y muestras.-</u>

Llamamos población (estadística) al grupo de individuos bajo estudio sobre el que deseamos hacer alguna inferencia, o sea al conjunto de objetos, mediciones u observaciones del cual tomamos nuestra muestra.

Una población puede ser finita o infinita, dependiendo de su tamaño. El tamaño de la población o sea el número total de los individuos que la conforman se lo denota con la letra **N**.
Ejemplos de poblaciones son:

- 20 camarones en una pecera.
- Todos los camarones de una piscina.
- Todos los camarones posibles a ser cultivados bajo cierto tratamiento.
- Todos los camarones del mundo.

Como se ve los límites de la población van a depender de como la definamos nosotros acorde con nuestras necesidades. Por esto antes de empezar cualquier proceso estadístico es necesario <u>Definir claramente la población o poblaciones bajo estudio</u>.

Conociendo la distribución de frecuencias de alguna característica (variable) de la población, es posible describirla por medio de una función de densidad, la cual a su vez vendrá caracterizada por ciertos parámetros (mas adelante veremos estos conceptos).

El problema principal de donde se originó la estadística es que al ser la población completa generalmente muy grande para ser estudiada en su totalidad, resulta más conveniente estudiar solo un subconjunto de dicha población y decir que este subconunto (muestra) representa mas o menos fielmente (representativo) a la población total.

Sea una variable aleatoria dada (X); los valores de esta variable aleatoria (X1, X2,...Xn) forman una muestra de la variable X, si ellas son independientes y siguen la misma distribución de X, en otras palabras, si representa fielmente a X.
Esto quiere decir que nosotros estamos partiendo de la suposición de que una muestra es una porción de una población que representa fielmente a la misma. O sea que no tomamos en cuenta los muestreos mal realizados. Por esto nosotros debemos de tomar todas las consideraciones técnicas prácticas necesarias para que esto se cumpla.

El tamaño de la muestra se lo denota como **n**.

Al igual que la población, la muestra debe de definirse correctamente antes de empezar el estudio. Así mismo:

- 20 camarones en una pecera.
- Todos los camarones de una piscina.
- Todos los camarones posibles a ser cultivados bajo cierto tratamiento.
- Todos los camarones del mundo.

pueden representar una muestra de una población mayor. La muestra y la población, así como sus límites lo fija el investigador de acuerdo a sus necesidades.

El objetivo de los muestreos es obtener información sobre las distribuciones de frecuencia de la población (distribución de probabilidad) o más preciso de los parámetros poblacionales que describen dicha distribución de probabilidad.

Estadísticos y Parámetros

La mayoría de las investigaciones estadísticas se proponen generalizar a partir de la información contenida en muestras aleatorias acerca de la población de donde fueron obtenidas.

En general, trataremos de hacer inferencias sobre los **parámetros** de las poblaciones (por ejemplo la media $\mu$ o la varianza $\sigma^2$). Que describen a la población. Estos generalmente se los denota con letras griegas.($\mu$,$\nu$, $\lambda$,$\sigma$, $\rho$, etc.). Entonces definimos **parámetros** como ciertas medidas que describen a la población. A los parámetros en general los podemos definir como $\theta$.

Para efectuar tales inferencias utilizaremos **estadísticos muestreales** o **estimadoes** como el promedio o media aritmética (x) y la varianza muestreal ($s^2$); es decir cantidades calculadas con base en datos u observaciones de la muestra.
Definimos **estadístico** como una medida que describe a la muestra. A los estadísticos en general los podemos definir como $\theta_n$.

Es importante aclarar la diferencia entre estadístico y parámetro, por que esa es una de las bases de la estadística. A pesar de que como veremos mas adelante, ciertos estadísticos como el promedio se los usa para representar a parámetros como la media, la probabilidad de que sean exactamente iguales es en realidad 0.
Veamos por ejemplo el promedio, este es un estadístico (estimador), ya que es una función de las observaciones de la muestra. Sin embargo, nótese que el promedio es una variable aleatoria y tiene una distribución de probabilidad o distribución de muestreo que depende del mecanismo de muestreo. Algunos de los valores que puede tomar el promedio estarán cerca de $\mu$, y otros pueden estar bastante alejados de ella, tanto para arriba como hacia abajo. Si nosotros tomamos varias muestras y calculamos el promedio, desearíamos que en promedio, el promedio nos dará valores concentrados cercanos a $\mu$, y que estén lo mas cercano a $\mu$. Entonces lo que decimos es que queremos seleccionar un estimador y un plan de muestreo que:

1. Nos asegure que la esperanza de el estimador sea la media ($E(\theta_0) = \mu$)
2. La varianza del estimador tenga la menor varianza ($\sigma(\theta_0) \rightarrow$ sea baja)

El estadístico que posee la primera característica se llama **insesgado**, el que que tiene la segunda se llama estadístico **eficiente**. De dos estadísticos $\theta_1$ y $\theta_2$, el que tenga menor varianza será el mas eficiente.

Aunque la distribución de probabilidad del promedio dependerá en gran parte del mecanismo de muestreo y del tamaño de la muestra y de la población, la distribuciónn muestreal tiende a presentar una distribución Normal. Esto es especialmente cierto si el tamaño de la muestra (n) es grande.

Una vez que sabemos que estadístico $\theta_0$ estamos usando, y coocemos su distribución de probabilidad, podemos evaluar su error de estimación. Definimos error de estimación como el valor absoluto de la diferencia entre el estadístico y el parámetro ($E=|\theta_0 - \theta|$). No se puede decir exactamente cual es este error, ya que desconocemos el parámetro ($\theta$), pero al menos podemos encontrar unos límites entre los cuales podamos decir que existe una probabilidad de que se encuentre el parámetro $\theta$ : $P(|\theta_0 - \theta|) \leq 1-\alpha$. El valor de $\theta_0$ + E se conoce como limite de confianza inferior, y $\theta_0$ - E como limite inferior de confianza.

Estadísticos de centralización.-

Podemos definir al **parámetro media poblacional** $\mu$ como la media aritmética de todos los individuos de nuestra población, y representa la esperanza matemática de nuestra variable aleatoria:

$$\mu = \frac{1}{N}\sum_{1}^{N} x_i$$

Este parámetro no lo conocemos, y no lo conoceremos nunca a no ser que muestreáramos la población completa. Es por esto que para estimar este parámetro utilizamos el **estadístico promedio o media muestreal**.

Nuestro promedio x va a estar dado por la media aritmética de los datos de nuestra muestra. Su fórmula es:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Al ser la media el valor de la esperanza matemática de los promedios, esta puede calcularse también de la siguiente forma:

$$\mu = \frac{n_j}{N}\sum_{1}^{k} \mu_j$$

En donde j es el j-esimo grupo de un total de k grupos, $n_j$ es el número de individuos en el j-esimo grupo y $\mu_j$ es la media del j-esimo grupo
Así mismo, se puede calcular un promedio ponderado x^:

$$\hat{x} = \frac{n_j}{n} \sum_{1}^{k} \overline{x}_j$$

Otros estadísticos de centralización son la moda y la mediana.

La moda corresponde a la marca de clase del intervalo de clases con mayor frecuencia (esto lo veremos mas claramente al hablar de distribuciones de frecuencias). En términos aproximados, es el valor que mas encontramos en nuestro muestreo.
La mediana correspondería al valor del dato que se encuentra más cercano a la mitad si ordenáramos nuestros datos, o al valor del dato que tiene igual número de datos mayores de el que menores de él.

La mediana viene dada por el valor del dato número (n+1)/2 cuando n es impar y por la media del dato # (n/2) y el dato # (n/2 +1) cuando n es par.

Estadísticos de dispersión.-

Las medidas de centralización nos dan una idea de hacia dónde están distribuidos nuestros datos, pero no de cómo están distribuidos. Es más, la probabilidad de que en un muestreo encontremos un individuo con un valor igual a la media es bajísima (por definición es 0). Por ejemplo, la media de los posibles valores que puede obtener el lanzamiento de un dado es (1/6 +2/6+3/6+4/6+5/6+6/6 = 3.5), y este valor es imposible que salga en un lanzamiento de dados. Un ejemplo de esto es un chiste que dice (mal) que un estadístico es un tipo que tiene los pies en un horno y la cabeza en la refrigeradora y dice “en promedio estoy bien” ( en realidad diría en promedio estoy bien, pero la varianza es insoportable).
Además, podemos tener dos poblaciones que aunque tengan igual media, la dispersión de los datos sea totalmente distinta, por lo cual las poblaciones son en realidad distintas. Para esto tenemos las medidas de dispersión.

El **parámetro varianza poblacional** $\sigma^2$ mide el promedio de los cuadrados de las desviaciones de todos los valores de una variable de una población con respecto a la media poblacional. Este parámetro equivale a la siguiente fórmula:

$$\sigma^2 = \frac{\sum(x_i - \mu)^2}{N}$$

Esto nos indica que tan lejos de la media se encuentran en promedio los datos (nótese que $x_i$-$\mu$ es la distancia a la que cada punto se encuentra de la media). Se lo eleva al cuadrado porque de no hacerlo las distancias positivas se anularían con las negativas y el resultado sería siempre 0.

La varianza empírica $s^2$ es el estadístico mediante el cual hacemos estimaciones de nuestro parámetro varianza poblacional. Debido a que la varianza empírica sería un estimador sesgado de la varianza poblacional si la dividiéramos para n, el cálculo de la misma va a estar dada por:

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n - 1}$$

Nótese que a medida que el tamaño de la muestra (n) aumenta, la diferencia entre $\sigma^2$ y $s^2$ disminuye.

La desviación típica o desviación estándar ($\sigma$ o s), no es otra cosa que la raíz cuadrada positiva de la varianza.

En las calculadoras con funciones estadísticas, $\sigma$ se denota como $\sigma_n$ y s como $\sigma_{n-1}$.
En Lotus v.2.3 o mas, $\sigma$ es @std(RANGO) y s @sstd(RANGO).
En las versiones anteriores de Lotus había que hacer una corrección:
s2 = @VAR(RANGO)*@COUNT(RANGO)/(@COUNT(RANGO)-1)
s = @STD(RANGO)*@SQRT(@COUNT(RANGO)/(@COUNT(RANGO)-1))
En otras hojas electrónicas refiérase al manual o a la ayuda del programa antes de usar las fórmulas indistintamente.

El rango es la diferencia entre el valor de nuestro mayor dato y el valor de nuestro menor dato.

**Probabilidades y Distribuciones.-**

Probabilidad

La base sobre la cual se fundamenta toda la teoría estadística es la probabilidad, es por esto que le estamos dando un gran énfasis a este capítulo en este curso. El entendimiento de la probabilidad permitirá concatenar los diferentes procesos que se utilizarán, y permitirán resolver situaciones nuevas analizándolas desde este punto de vista. Más importante que entender los cálculos que vamos a realizar en determinada situación o como buscar en las tablas que en determinada situación vayamos a utilizar (ya que con una computadora se lo puede hacer casi instantáneamente), es entender el concepto mismo de probabilidad, poder entender las relaciones que se dan en determinada circunstancia entre una variable en una población y la distribución de dicha variable en la tabla (o curva) y poder situarnos en dicha curva y usarla como guía para determinar la probabilidad de ocurrencia de un suceso (o el rango de sucesos que pueden suceder con determinada probabilidad)

Teoría de probabilidades.-

Eventos que son comunes o improbables son aquellos cuya probabilidad de ocurrencia son grandes o pequeñas, respectivamente.
Sin darnos cuenta, nosotros calculamos "al ojo" la probabilidad de todas los sucesos que nos rodean; así, determinamos que tan "común" o "raras" son ciertos acontecimientos.
Por ejemplo, en Esmeraldas no es "común" encontrar un nativo rubio y ojos azules pero en Suecia si. Esto lo sabemos en base a "muestras" de Suecos y Esmeraldeños que hemos visto, sin necesidad de ver todos los esmeraldeños y todos los suecos.
El problema de este método al "ojímetro" es que carecemos de un término preciso para describir la probabilidad.

Los estadísticos reemplazan las palabras informativas pero imprecisas como "con dificultad", "pudo" o "casi con seguridad" por un número que va de 0 a 1, lo cual indica de forma precisa que tan probable o improbable es el evento.
Lógicamente, haciendo inferencias a partir de muestras sobre una población, es decir de una parte sobre el todo, no podemos esperar llegar siempre a resultados correctos, pero la estadística nos ofrece procedimientos que nos permiten saber cuántas veces acertamos "en promedio". Tales enunciados se conocen como enunciados probabilísticos.

Llamamos espacio muestreal al conjunto universal de una población o a todos los valores probables que nuestra variable aleatoria puede tomar. Ejemplos.- Todas las formas en que podemos sacar 4 bolas de una funda que contenga 8 bolas rojas y 2 blancas, de cuantas formas puede caer un dado, todas las posibles supervivencias que podamos obtener en un cultivo, todos los posibles climas que puedan haber en un día determinado, todos los tamaños o pesos que pueda tener una especie en cultivo, etc.

Matemáticamente, si un evento puede ocurrir de **N** maneras mutuamente exclusivas e igualmente posibles, y si **n** de ellas tienen una característica **E**, entonces, la posibilidad de ocurrencia de E es la fracción **n**/**N** y se indica por:

$$\mathrm{P}(E) = \frac{n}{N}$$

O sea, que la probabilidad de que ocurra un evento determinado (éxito) no es más que la razón de el número de éxitos divididos para el tamaño total del espacio muestreal (éxitos + fracasos).
La definición de "éxito" o "fracaso" no tiene nada que ver con la bondad del suceso y se lo asigna de forma arbitraria de acuerdo a nuestras necesidades, por ejemplo puede considerarse un "éxito" el número de niños que se enferman en un año, o la cantidad de larvas que se mueran durante una aclimatación.

En general, para sucesos en los cuales el tamaño de el espacio muestreal nos sea desconocido o infinito, cuando no podemos saber la cantidad total de éxitos o cuando todas las maneras en que pueda ocurrir el suceso no sean igualmente "posibles", recurriremos al muestreo, esto es, definiremos la probabilidad como "la proporción de veces que eventos de la misma clase ocurren al repetir muchas veces el experimento", y esta es la definición que más usaremos en nuestro curso.
Por ejemplo, la probabilidad de que un carro sea robado en Guayaquil en un período de tiempo puede ser calculada empíricamente en función al número de carros robados en dicho período de tiempo y al número de carros en Guayaquil. Esto es lo que hacen las aseguradoras, y utilizan esta probabilidad para calcular el valor esperado a pagar, le aumentan los costos y la utilidad y de ahí determinan cual es la prima a pagar.
O la probabilidad de que en cierta camaronera una corrida a 130.000 Pl/Ha alcance 15 gr. en 140 días es calculada en base al número de veces que se ha logrado esto en condiciones similares, o la posibilidad de que llueva en un día determinado por la cantidad de veces que llueve en condiciones atmosféricas similares.

En todos estos casos hay una probabilidad (1-p) de no ocurrencia, entonces si hay una probabilidad del 0.99 (99%) de que algo ocurra, significa que hay también una probabilidad de 0.01 (1%) de que no ocurra. en general, se considera que probabilidades de ocurrencias de menos del 5% (0.05) son “poco comunes”. Pero debemos de tomar en cuenta de que si algo es “poco probable” de que ocurra no significa que no va a ocurrir”.

Teoremas básicos.-

- La probabilidad de un evento cualquiera va a estar en el rango de cero a uno. Esto quiere decir que no existen probabilidades negativas ni mayores de 100%

  $0 \leq P(E) \geq 1$

- La suma de la probabilidad de ocurrencia de un evento mas la probabilidad de no ocurrencia del mismo es igual a uno.

  $P(E) + P(\neg E) = 1$

- Para dos eventos cualesquiera A y B, la probabilidad de que ocurra A o B viene dado, por la probabilidad de que ocurra A, mas la probabilidad de que ocurra B, menos la probabilidad de que ocurran ambos.

  $P(A \text{ o } B) = P(A) + P(B) - P(AB)$

  Si los eventos son mutuamente exclusivos, P(AB) será 0 y la probabilidad de ocurrencia de ambos será : $P(A \text{ o } B) = P(A) + P(B)$

Valor esperado.-

Llamamos **valor esperado** al valor probable que podemos obtener al repetir cierto evento. Este valor va a estar asociado a la probabilidad de ocurrencia de dichos eventos dicho evento y al valor que tomará la variable en los distintos eventos.

Por ejemplo, si decimos que la probabilidad de que ganemos al apostar a un número en la ruleta es 1/37 = 0.27, y que si apostamos 1,000,000 sucres y ganamos obtendremos 35,000,000 sucres y si perdemos 1,000,000, entonces la esperanza de ganar en la ruleta será:

E(ganancia) = P(ganar)*Valor a ganar + P(perder)* Valor a Perder
o
E(ganancia) = 1/37 * 35,000,000 + 36/37 (-1,000,000)
E(ganancia) = 945,946 - 972,973 = -27,027

O en otras palabras, si jugamos a la ruleta, apostando toda la noche a un número 1,000,000 sucres, la esperanza que tenemos es de perder “en promedio” 27,000 sucres cada vez.

Esto tiene bastantes aplicaciones prácticas, de las cuales veremos algunos casos.

Distribución de probabilidad.-

Una distribución o densidad de probabilidad de una variable aleatoria x es la función de distribución de la probabilidad de dicha variable o, en otras palabras, la probabilidad de que dicha variable tome ciertos valores.

Observemos por ejemplo la distribución de probabilidad normal, en ella, el área de la curva entre 2 puntos cualesquiera representa la probabilidad de que ocurra un suceso entre esos dos puntos.

Las distribuciones de probabilidad pueden ser discretas o continuas, de acuerdo con el tipo de variable al cual representen.

Hay infinidad de distribuciones de densidad, una para cada población, pero se han definido ciertas distribuciones "modelo" más comunes como la Normal, binomial, Ji-cuadrado, "t" de Student, F de Fisher; a las cuales podemos aproximar estas distribuciones particulares.

Empezaremos con revisar la distribución Normal.

Distribución Normal
Como Uds. se pueden dar cuenta, hemos otorgado a esta distribución un capítulo entero en el manual, y esto se debe a que no es exagerado decir que es la distribución mas importante que estudiaremos.

Esta Distribución fue descubierta en 1733 por el francés Moiure, y fue descrita también por Laplace y por Gauss, siendo el nombre de este último también sinónimo de la forma gráfica de esta distribución.

Como ven en la gráfica, esta distribución tiene forma de campana y presenta ciertas características importantes:

- El área debajo de la curva entre 2 puntos dados representa la probabilidad de que ocurra un hecho entre esos dos puntos;
- Su dominio va de menos infinito a más infinito;
- Es simétrica con respecto a su media;
- Tiene dos colas y es asintótica al eje x por ambos lados;
- El valor del área debajo de toda la curva es igual a 1;
- El centro de la curva está representado por la media poblacional ($\mu$).
- Para cualquier curva normal, el área de $-\sigma$ a $+\sigma$ es igual a 0.6827; de $-2\sigma$ a $+2\sigma$ de 0,9545 y de $-3\sigma$ a $+3\sigma$ de 0,9973;
- La distribución muestreal de varios estadísticos, tales como la media, tienen una distribución aproximadamente normal e independiente de la configuración de la población.

La importancia práctica de esta distribución teórica de probabilidad estriba en que muchos fenómenos biológicos presentan datos distribuidos de manera tan suficientemente Normal que su distribución es la base de gran parte de la teoría estadística usada por los biólogos.

<u>Distribución Normal tipificada.-</u>

Llamamos distribución normal tipificada a la distribución especial que representa a todas las variables aleatorias normales y que es la distribución de otra variable normal llamada Z.
En donde Z va a ser igual a:

$$Z = \frac{x - \mu}{\sigma}$$

Y se la conoce como variable aleatoria estandarizada.

Esta función se caracteriza por tener media igual a cero (0) y desviación tipificada igual a uno (1).

Esta distribución Normal tipificada o estandarizada representa a todas las distribuciones Normales, sea cual sea su media y su varianza, Teniendo la misma densidad de probabilidad, si medimos las desviaciones de su media en base a s. Por lo tanto los valores obtenidos de la tabla Normal son válidos par todas las distribuciones Normal de media = $\mu$ y varianza =$\sigma^2$.

<u>Unos ejemplos del uso de la tabla:</u>
Obtenga la probabilidad de que Z obtenga los siguientes valores:

a.-

$P(0 \leq Z \leq 1.17)$

Buscamos en la columna derecha de la tabla el valor 1.1, y en la primera fila el valor 7 (correspondiente al decimal 0.07), interceptamos ambos valores obteniendo el valor de 0.3790, que es el valor que buscábamos:

$$P(0 \leq Z \leq 1.17) = 0.379$$

b.-

$P(Z \leq 1.17)$

Esto lo podemos escribir de la siguiente forma también:

$$P(0<Z<1.17) + P(Z \leq 0)$$

El primer término lo conocemos, por que lo resolvimos en el literal a.
Para el segundo término sabemos que la distribución normal es simétrica y que su área total es igual a 1, por lo tanto el área que hay de $-\infty$ a 0 ($P(Z \leq 0)$) es igual a 1/2 = 0.5.

Por lo que el valor que buscábamos estará dado por:

$$P(Z \leq 1.17) = 0.379 + 0.5 = 0.879$$

c.-

$P(Z \geq 1.17)$

Sabiendo que el área total bajo toda la curva Normal de $-\infty$ a $+\infty$ es igual a 1, y conociendo el valor del área de $-\infty$ a 1.17, el valor del área de 1.17 a $+\infty$ va a ser :

$$1 - P(Z \leq 1.17) = 1 - 0.879 = 0.121$$

d.-

$P(Z \leq -1.17)$

Como estamos tratando con una curva simétrica, este valor será el mismo que el del literal c:

$$P(Z \leq -1.17) = P(Z \geq 1.17) = 0.121$$

e.-

$P(0.42 \leq Z \leq 1.17)$

En este caso se trata de un intervalo cuyos límites se encuentran en la parte positiva de la curva.
El valor del área entre ambos puntos será igual al valor del área de 0 al punto superior menos el valor del área de 0 al punto inferior, esto es:

$$P(0 \leq Z \leq 1.17) - P(0 \leq Z \leq 0.42) = 0.379 - 0.1628 = 0.2162$$

f.-

$P(-1.17 \leq Z \leq -0.42)$

En este caso se trata de un intervalo cuyos límites se encuentran en la parte negativa de la curva.
Al tratarse de una curva simétrica, el valor del área será igual al del literal e.

$$P(-1.17 \leq Z \leq -0.42) = P(0.42 \leq Z \leq 1.17) = 0.2162$$

g.-

$P(-1.17 \leq Z \leq 0.42)$

En este caso se trata de un intervalo cuyos límites se encuentran a ambos lados del centro de la curva.
El área total será igual al área de 0 a cada uno de los límites del intervalo:

$$P(-1.17 \leq Z \leq 0) + P(0 \leq Z \leq 0.42) = 0.379 + 0.1628 = 0.5418$$

h.-

$P(|Z| \geq 1.17)$

En este caso se trata de determinar el área de $-\infty$ a -1.17 y de 1.17 a $+\infty$. Como la curva es simétrica, simplemente multiplicamos el valor de $P(Z \geq 1.17)$ del literal c por 2:

$$P(|Z| \geq 1.17) = 2 \times P(Z \geq 1.17) = 2 \times 0.121 = 0.242$$

i.-

$P(|Z| \leq 1.17)$

En este caso el valor del área va a estar dado por 1 menos el valor del literal h, ya que el valor total del área es igual a 1:

$$P(|Z| \leq 1.17) = 1 - P(|Z| \geq 1.17) = 1 - 0.242 = 0.758$$

Otro caso diferente para el cual podemos utilizar la tabla es para encontrar el valor de Z después del cual se encuentra un $\alpha \times 100$ % del área de la curva.
Esto equivale a decir buscar el valor de Z cuya probabilidad de ser mayor sea $100 \times \alpha$ %, o en su defecto que su probabilidad de ser menor sea de $(1-\alpha)\times100$ %

Por ejemplo:

a.- Hallar el valor de Z después del cual se encuentra el 5% del área de la curva:

Esto corresponde a un valor de $\alpha = 0.05$

Esto equivale a decir buscar el valor de Z tal que:

$$P(Z \geq x) = 0.05$$

Como en la tabla tenemos el área de 0 a Z, buscamos en el cuerpo de la tabla el valor de: 0.5 - 0.05 = 0.45

Vemos que este valor se encontraría en la fila correspondiente a 1.6, entre los valores de las columnas 4 (0.4495) y 5 (0.4505), por lo que al interpolar sacamos que se encuentra en la columna 4.5, entonces el valor de Z sería igual a 1.645.

$$P(Z \geq 1.645) = 0.05$$

b.- Hallar el valor de Z tal que el área sobre el mas el área bajo -Z sea igual a 0.05.

Esto equivale a decir buscar el valor de Z tal que:

$$P(|Z| \geq x) = 0.05$$

Como estamos tratando con una curva simétrica podemos decir que esto es igual a:

$P(Z \geq x) = 0.05/2 = 0.025$

Como en la tabla tenemos el área de 0 a Z, buscamos en el cuerpo de la tabla el valor de: 0.5 - 0.025 = 0.475

Vemos que este valor se encuentra en la fila correspondiente a 1.96, y en la columna correspondiente a 6, por lo que entonces el valor de Z sería igual a 1.96.

$P(Z \geq 1.96) = 0.25$

ó :

$P(|Z| \geq 1.96) = 0.05$

Bueno, todo esto estaría bien si se tratara de una curva Normal tipificada ($\mu=0$; $\sigma=1$), pero que pasaría si quisiéramos usarlo en una población natural con una media : $\mu \neq 0$ y desviación estándar: $\sigma \neq 1$. Bueno, no hay problema, solo tenemos que tipificar el valor de x en nuestra distribución Normal ($\mu,\sigma$) mediante la fórmula:

$$Z = \frac{x - \mu}{\sigma}$$

y procedemos a buscar la probabilidad para este valor determinado. Como podemos ver en la fórmula, Z no es ni mas ni menos que el número de desviaciones estándares de distancia a la que se encuentra el valor x de la media m.

Por ejemplo:

Encontrar la probabilidad que al muestrear una piscina que contenga una población Normal con $\mu=5$ y $\sigma^2=4$ encontremos un valor mayor que 7.78.

Como la varianza es $\sigma^2=4$, entonces $\sigma = 2$.

calculamos el valor de Z:

$$Z = \frac{7.78 - 5}{2} = 1.39$$

y luego calculamos la probabilidad de que Z sea mayor a este valor en la tabla:

$P(Z \geq 1.39) = 0.5\text{-}0.4177=0.0823$

Al aplicar pruebas de hipótesis esto generalmente no es necesario realizarlo, ya que los cálculos están hechos en base a la distribución Normal tipificada.

Distribución Derivada.-
Cuando muestreamos repetidamente una población, podemos obtener una distribución de sus medias muestreales llamada distribución derivada. La media de una población de promedios de n observaciones es igual a la media de la población, y su varianza es igual a 1 n-esimo de la varianza poblacional ($\sigma^2/n$), o sea:

$$\mu x = \mu \qquad\qquad \sigma^2 x = \sigma^2/n$$

Por ejemplo: encontrar la probabilidad que al sacar una muestra de tamaño n=16 de una población con $\mu$=10 y $\sigma^2$=4 encontremos un promedio mayor o igual a 11.

El promedio x es una muestra tomada de una población normal derivada con:

$$\mu x = \mu = 10$$

y:

$$\sigma^2 x = \sigma^2/n = 4/16 = 1/4 \ ; \quad \sigma x = 1/2$$

Entonces buscamos el valor de Z, o sea la distancia a la que nuestro promedio se encuentra de la media:

$$Z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{11 - 10}{1/2} = 2$$

Buscando en la tabla encontramos que:

$P(Z \geq 2) = 0.228$

Por lo que podemos decir que la probabilidad de sacar un muestreo de n = 16 y x $\geq$ 11 es de 2.28%, lo cual se considera "poco usual".

En realidad esto no se aplica tan bien para muestras < 30, pero esto era solo un ejemplo. más adelante veremos que pasa con las muestras pequeñas.

Entre las aplicaciones que veremos de la distribución Normal se encuentran:

- Estimaciones de intervalos de confianza para la media.
- Pruebas de hipótesis con respecto a medias.
- Aproximaciones a otras distribuciones de probabilidad.

Distribución "t" de Student.-

Desarrollada con base en distribuciones de frecuencia empíricas en 1908 por William Gosset, conocido por el alias de "Student". Un cervecero - estadístico que encontraba ciertas dificultadas al usar la distribución Normal en muestras pequeñas. Esta misma tabla ya había sido calculada matemáticamente en 1875 por un astrónomo alemán, el cual sin embargo no le había encontrado utilidad práctica.

El problema reside en que la distribución muestreal de la media se ajusta muy bien a la distribución Normal cuando se conoce $\sigma$. Si **n** es grande, esto no presenta ningún problema, aun cuando $\sigma$ sea desconocida, por lo que en este caso es razonable sustituirla por ***s***. Sin embargo, en el caso de usar valores de **n** < 30, o sea en el caso de pequeñas muestras, esto no funciona tan bien.

Definiendo el estadístico ***t***:

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$

Se puede probar que siendo x el promedio de una muestra tomada de una población normal con media $\mu$ y varianza $\sigma^2$, el estadístico ***t*** es el valor de una variable aleatoria con distribución "t" de Student y parámetro $\nu$ (grados de libertad) = n-1.

Como puede apreciarse en la figura, la distribución "t" es muy similar a la distribución normal, y entre sus características tenemos:

- Tiene media igual 0, es asintótica al eje x y su dominio va de $-\infty$ a $+\infty$;
- El área bajo la curva desde $-\infty$ a $+\infty$ es igual a 1;
- Al igual que la distribución Normal estándar, esta distribución tiene media 0, pero su varianza depende del parámetro $\nu$, denominado grados de libertad;
- La varianza de la distribución "t" excede a uno, pero se aproxima a ese número cuando $n \Rightarrow \infty$,
- Al aumentar el valor de n, la distribución "t" de Student se aproxima a la distribución Normal, es más, para tamaños muestreales de 30 ó más, la distribución Normal ofrece una excelente aproximación a la distribución "t".

Entre las aplicaciones de esta distribución tenemos la estimación de intervalos de confianza para medias a partir de muestras pequeñas y las pruebas de hipótesis basadas en muestras < 30.

En la tabla correspondiente se encuentran los valores de $t_\alpha$ a la derecha de los cuales se encuentra un (100 x $\alpha$)% del área de la curva.

Para buscar el valor de $t_a$ para n = n-1 en la tabla, primero localizamos la columna del correspondiente valor de $\alpha$ y la fila correspondiente al valor de $\nu$. La intersección de la fila y la columna nos dará el valor de $t_\alpha$.

Ji-cuadrado.-
La distribución Ji-cuadrado es una función de densidad de probabilidad que sigue aproximadamente una distribución gamma con $\alpha = n/2$ y $\beta = 2$, y que representa la distribución muestreal de la varianza.

Definimos el estadístico Ji-cuadrado ($\chi^2$) como:

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

Entre sus características tenemos:

- Es asimétrica y asintótica al eje x por la derecha;
- Su dominio va de 0 a $+\infty$
- El área bajo la curva desde 0 a $+\infty$ es igual a 1;
- Tiene parámetro $\nu$ = n-1;
- Al incrementarse n se aproxima a la distribución normal; y,
- Representa la distribución muestreal de la varianza.

Entre las aplicaciones de esta distribución tenemos la determinación de intervalos de confianza para varianzas, las pruebas de hipótesis para una varianza, el ajuste de datos a una distribución dada conocida y las pruebas de independencia.

En la tabla correspondiente se dan algunos valores de $\chi^2$ para varios valores de $\nu$, donde $\chi^2$ es tal que el área bajo la curva, a su derecha, es igual a $\alpha$. En esta tabla, la columna del lado izquierdo contiene los valores de n, el primer renglón consta de áreas a en la cola del lado derecho de la distribución Ji-cuadrada y la tabla propiamente dicha la constituyen los valores de $\chi^2$.

"F" de Fisher - Schnedecor.-
Otra distribución de probabilidad muy usada es la "F" de Fisher - Schnedecor, la cual está relacionada con la distribución beta, y representa la distribución muestreal de la razón de dos varianzas. Es decir que se obtiene de la razón de dos distribuciones Ji-cuadrado.

Definimos el estadístico F como:

$$F = \frac{s^2_1}{s^2_2}$$

El cual es el valor de una variable aleatoria que tiene distribución F con parámetros $\nu_1=n_1-1$ y $\nu_2=n_2-1$.

Su distribución gráfica la podemos apreciar en la figura.

Entre sus propiedades tenemos:

- Es asimétrica, y asintótica al eje x por el lado derecho;
- Su dominio va de 0 a $+\infty$;
- El área bajo la curva desde 0 a $+\infty$ es igual a 1; y,
- Tiene parámetros $\nu_1=n_1-1$ y $\nu_2=n_2-1$.

La tabla correspondiente contiene los valores de F de $\alpha$ = .01 y .05 para varias combinaciones de $\nu_1$ y $\nu_2$.

Entre las aplicaciones de esta distribución están las pruebas de hipótesis entre 2 varianzas y los análisis de varianza y covarianza.

Distribución binomial.-
Muchos ensayos poseen solo dos resultados posibles, por ejemplo un animal sobrevive o no a cierto tratamiento, posee o no cierta característica. Estos fenómenos presentan generalmente una distribución de densidad asociada con la distribución **binomial**.

Interpolación de valores en tablas estadísticas.-
Veremos 2 formas de interpolación, la lineal y la armónica.

La interpolación lineal se la usa cuando el valor **x** para el cual queremos hallar el valor **G** de la función no se encuentra en la tabla pero se encuentra entre dos valores **x1** y **x2** dados, los cuales tienen sus respectivos valores **G1** y **G2** de la función. Aquí, el valor de **G** vendrá dado por:

$$G = G_1 + \frac{x - x_1}{x_2 - x_1}(G_2 - G_1)$$

Por ejemplo, si queremos el valor de $t_{0.01}$ para n=33 grados de libertad, buscamos el valor de :

x1 = n1 = 30, que es el valor inmediatamente menor que x, y cuyo valor en la tabla de G1= 2.46.

x2 = n2 = 40, que es el valor inmediatamente mayor que x, y cuyo valor en la tabla de G1= 2.42.

aplicando la fórmula tenemos:

$$G = 2.46 + \frac{33 - 30}{40 - 30}(2.42 - 2.46)$$

G = 2.46 + 3/10 (-0.04) = 2.448

entonces:

$t_{0.01\ (n=33)}$ = <u>2.448</u>

En algunas tablas el último valor es infinito, en estos casos se recomienda aplicar una interpolación armónica, en donde reemplazamos x, x1 y x2 por x2/x, x2/x1 y x2/x2 respectivamente, siendo x el valor para el cual deseamos encontrar G, x1 infinito y x2 el último valor numérico de la tabla. Después de esto utilizamos la misma formula de la interpolación lineal.

Por ejemplo si queremos el valor de $F_{0.05}$ para $n_1$=2 y $n_2$=150 grados de libertad tendremos:

x1 = ∞ x = 150 x2 = 120

x1'= 120/∞ =0 x'= 120/150 = 0.8 x2'= 120/120 = 1

G1 = F(2,∞)=3.00 G = F(2,150)= ? G2 = F(2,120)=3.07

Luego de lo cual aplicamos la fórmula anterior.

$$G = 3.00 + \frac{0.8 - 0}{1 - 0}(3.07 - 3.00)$$

G = 3.00 + 0.8 (0.07) = 3.056

$$F_{0.01}(2,150) = \underline{3.056}$$

**Estadística Descriptiva.**

Los datos estadísticos, obtenidos de muestras, experimentos o cualquier colección de mediciones, a menudo son tan numerosos que carecen de utilidad a menos que sean condensados o reducidos a una forma más adecuada. Por ello, en esta sección nos ocuparemos del agrupamiento de datos, así como de ciertos estadísticos o medidas que representarán el significado general de nuestros datos.

Distribución de Frecuencias

Vamos a ver a continuación la distribución de frecuencias.

La distribución de frecuencias es una operación mediante la cual dividimos un conjunto de datos en un número de clases apropiadas, mostrando también el número de elementos en cada clase.

Ordenando los datos en una distribución de frecuencias, se va a perder cierta cantidad de información, pero esto se compensa con lo que ganamos en claridad.

La primera etapa en la construcción de una distribución de frecuencias consiste en decidir cuántas clases utilizar, y elegir los límites de cada clase. En general, el número de clases dependerá del número y rango de los datos. Matemáticamente, el número de intervalos (k) viene dado por la siguiente fórmula:

$$k = 1 + \frac{10}{3} \ln n$$

aunque siempre hay que ver qué tan bien representa esto la veracidad de los datos. Empíricamente, se recomienda usar un número de intervalos no menor que 5 o mayor que 15.

Llamamos intervalo de representación al intervalo donde se representan los datos.

Intervalo real son los verdaderos límites del intervalo de representación, y viene dado por el punto medio entre los límites de dos intervalos de representación consecutivos.

Definimos la marca de clase como el punto medio entre el límite superior y el inferior de un intervalo de representación.

La frecuencia es la cantidad de ocurrencias de datos dentro de un intervalo de representación.

La frecuencia relativa es la relación entre la frecuencia de un intervalo y la frecuencia total expresada en porcentaje.

La frecuencia acumulada y acumulada relativa es la sumatoria del número de ocurrencias o porcentajes de todos los intervalos menores o iguales al presente.
Por ejemplo, con los siguientes datos de longitud cefálica en:

*Tilapia nilótica:*

| | | | | | | |
|---|---|---|---|---|---|---|
| 29.0 | 29.0 | 27.0 | 28.2 | 28.9 | 26.1 | 28.4 |
| 29.4 | 28.2 | 26.0 | 29.5 | 27.4 | 25.3 | 26.0 |
| 27.0 | 27.6 | 26.0 | 29.7 | 29.8 | 26.3 | 28.5 |
| 30.2 | 27.0 | 28.0 | 31.0 | 25.6 | 33.4 | 28.0 |
| 29.0 | 28.0 | 29.5 | 26.4 | 27.3 | 29.3 | 26.0 |
| 28.0 | 26.0 | 29.5 | 29.5 | 29.4 | 26.6 | 26.4 |
| 28.0 | 27.7 | 28.1 | 27.6 | 26.8 | 27.0 | 29.3 |
| 28.0 | 27.0 | 31.0 | 27.0 | 27.0 | 28.9 | 29.3 |

Construimos una tabla parecida a la siguiente:

| **¡Error! Marcador no definido.** Intervalo Represent. | Interv. Real | Frecuen. | Frecuen. Relativa | Frecuen. Acumulada | F. Acum. Relativa | Marca de Clase |
|---|---|---|---|---|---|---|
| 24 - 25 | 23.5- 25.5 | 2 | 3.57 % | 2 | 3.57 % | 24.5 |
| 26 - 27 | 25.5 - 27.5 | 19 | 33.93 % | 21 | 37.50 % | 26.5 |
| 28 - 29 | 27.5 - 29.5 | 29 | 51.79 % | 50 | 89.29 % | 28.5 |
| 30 - 31 | 29.5 - 31.5 | 5 | 8.93 % | 55 | 28.22 % | 30.5 |
| 32 - 33 | 31.5 - 33.5 | 1 | 1.78 % | 56 | 100.00 % | 32.5 |

Las propiedades de las distribuciones de frecuencias relacionadas con su forma se hacen más evidentes por medio de gráficos.

La forma más común de representar una distribución de frecuencia es el histograma, en el cual el área de los rectángulos representan las frecuencias de clase, y sus bases se extienden en las fronteras de los intervalos reales. En este tipo de gráfico, las marcas de clase están situadas en la mitad del rango del rectángulo. Mediante este gráfico podemos representar la frecuencia o la frecuencia relativa, pero no la frecuencia acumulada o acumulada relativa.

Otros gráficos similares a los histogramas son los diagramas de barras, aquí, las alturas y no las áreas representan las frecuencias de clase, y no se pretende fijar ninguna escala horizontal continua; en otras palabras, el ancho de las barras no interesa. Por esto se pueden graficar tanto las frecuencias absolutas o relativas, así como las acumuladas.

Otra forma de presentar las distribuciones de frecuencia en forma gráfica es el polígono de frecuencias. En él, las frecuencias de clases son graficadas sobre las marcas de clase y unidas mediante líneas rectas. Además, agregamos valores correspondientes a cero en los puntos límites de la distribución.

Con estos gráficos podemos representar indistintamente las frecuencias netas o acumuladas; sin embargo, cuando graficamos estas últimas, en vez de utilizar las marcas de clase como abscisas utilizamos el límite superior del intervalo real de frecuencia.

Para presentar datos de frecuencias relativas, el gráfico de sectores o "pie" se usa con mucha frecuencia. Este corresponde a un círculo dividido en varios sectores, correspondiendo cada uno a un intervalo, y en donde el área de cada sector es proporcional a la frecuencia relativa.

<u>Tipos de Muestreos .Estimación de Parámetros y errores. Tamaño de la Muestra</u>

La estimación de parámetros es parte importante de la estadística descriptiva. Esto nos sirve para describir poblaciones, como por ejemplo los resultados de alguna prueba.

En esencia, la estimación puntual se refiere a la elección de un estadístico calculado a partir de datos muestreales, respecto al cual tenemos alguna esperanza o seguridad de que esté "razonablemente cerca" del parámetro que ha de estimar.

Se recuerdan cuando vimos la diferencia entre parámetro y estadístico?, pues una estimación puntual no es mas que calcular un estadístico, y decir que este estadístico esta "razonablemente cerca" del parámetro poblacional.

Llamamos estimador insesgado a un estadístico cuyos valores promedios son iguales a los del parámetro que trata de estimar.

Como ejemplo de esto tenemos el promedio y la media, si tomamos todas las posibles muestras de una población y le calculamos el promedio a los promedios de cada uno de ellos, el resultado será el parámetro media poblacional.

De dos estadísticos dados, $\theta_1$ y $\theta_2$, podemos decir que el más eficiente estimador de $\theta$ es aquel cuya varianza de distribución muestreal es menor.

Para poblaciones normales, el estimador más eficiente de $\mu$ es el promedio (x).
En lo que respecta a la varianza poblacional, el estimador insesgado más eficiente es la varianza muestreal.
Para la varianza el problema estaría que si dividimos para n nos dará un estimador sesgado, pero si dividimos para n-1, se convierte en un estimador insesgado, por esto es que dejamos definida en la primera clase a la varianza muestreal $s^2$ como un estimador insesgado de $\sigma^2$.
A pesar de que la varianza $s^2$ es un estimador insesgado de $\sigma^2$, la desviación muestreal s, no lo es respecto de $\sigma$, aunque para muestras de gran tamaño el sesgo es pequeño y acostumbra a usárselo.
Del rango muestreal R, también se puede sacar un estimador insesgado de $\sigma$, la relación R/d2 para tamaños muestreales $\leq 5$ es mas eficiente que s para estimar $\sigma$. Los valores de d2 son los siguientes para distintos valores de n:

| *n* | *2* | *3* | *4* | *5* | *6* | *7* | *8* | *9* | *10* |
|---|---|---|---|---|---|---|---|---|---|
| $d_2$ | 1.128 | 1.693 | 2.059 | 2.326 | 2.534 | 2.704 | 2.847 | 2.970 | 3.078 |

Y, para proporciones, el estimador insesgado mas eficiente del parámetro proporción poblacional (p) es el estadístico proporción muestreal (x/n).

$$x / n = \frac{x}{n}$$

En donde x es el número de observaciones con un caracter determinado y n es el número total de observaciones (x + ¬x).

Sin embargo, debemos recordar que cuando empleamos una estadístico muestreal para estimar el parámetro de una población, la probabilidad de que la estimación sea en realidad igual al parámetro es prácticamente nula. Por esta razón es conveniente acompañar la estimación puntual con una afirmación de cuán cerca podemos razonablemente esperar que se encuentre la estimación (error de estimación) o podemos utilizar la estimación por intervalos, que no es mas que decir que existe un determinada probabilidad de que el parámetro esté dentro de ese intervalo.

La forma de estimarlos parámetros van a depender del tipo de muestreo que se realice. Ya que las probabilidades de obtener un determinado valor van a variar de acuerdo al tipo de muestreo.

¿Como podemos cual procedimiento usar y el número de observaciones a incluir en la muestra?. Esto depende decuanta información se quiera y se pueda conseguir. Debemos de especificar un límite para el error de esimación, esto es que debemos de especificar que $\theta$ y $\theta_0$ difieran en una cantidad menor que $\Delta$. 0 E $\leq \Delta$. debemos de especificar también una probabilidad (1-$\alpha$) que especifique la fracción de la veces que queremos que al muestrear repetidamente la población, el error de estimación sea menor que $\Delta$. O $P(E \leq \Delta) = 1-\alpha$.

Despues de obtener un límite específico con su probabilidad asociada, podemos comparar diferentes métodos de selección de la muestra para determinar cual procedimiento proporciona la precisión deseada al mínimo costo.

Dos factores influyen en la cantidad de información contenida en una muestra. El primero es el tamaño de la muestra, y el segundo es la cantidad de variación que hay entre los individuos de una población. La variación puede ser controlada usualmente por el método de muestreo. Para un tamaño de muestra fija, consideramos diversos muestreos, puesto que las observaciones cuetan dinero, un diseño que proporcione un estimador preciso con el menor tamaño de muestra produce un ahorro en el costo del experimentado.

Por el enfoque del curso vamos a ver nosotros solamente Muestreo totalmente aleatorio y Muestreo aleatorio estratificado. Pero si alguien desea ampliar sus conocimientos sobre muestreos le recomiendo el libro Elementos de Muestreo de Richard Scheaffer.

Muestreo Totalmente aleatorio

El diseño básico es el **muestreo totalmente aleatorio** o **muestreo irrestricto al azar**, y consiste en seleccionar un muestreo de n unidades muestreales o individuos de tal forma que cada muestra de tamaño n tenga la misma oportunidad de ser seleccionada. A la muestra obtenida de esta manera se la llama **muestra totalmente aleatoria** Este diseño de muestreo es tan bueno como cualquier otro siempre y cuando todos los individuos de la población tengan similares caracteristicas en cuanto a la información que nos interese y no exista otra variable que no permita separarla en grupos distintos entre ellos pero mas homogenos dentro de ellos que la población original.

Seleccionar una muestra totalmente aleatoria no es tan facil como parece, se puede usar el criterio del muestreador para seleccionar “aleatoriamente” la muestra, esto se llama muestreo casual. Una segunda técnica es seleccionar una muestra que consideremos representativa de la población. Estos muestreos están sujetos a sesgos por parte del entrevistador y conducen a estimadores cuyas distribuciones de probabilidad no pueden ser evaluadas, por lo tanto ninguna de ellas es totalmente aleatoria.

Estimación de media, Varianza, proporción y errores.

Para estimar la media poblacional $\mu$ en un muestreo totalmente aleatorio utilizaremos el promedio x :

$$\mu = \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

y su error de estimacion para poblaciones infinitas o muy grandes respecto a la muestra será:

$$E = Z_{(\frac{\alpha}{2})} \cdot \frac{\sigma}{\sqrt{n}}$$

Para poblaciones finitas, o en su defecto, cuando la muestra representa un porcentaje alto de la población, este será :

$$E = Z_{(\frac{\alpha}{2})} \cdot \frac{\sigma}{\sqrt{n}} \sqrt{\left(\frac{N-n}{N}\right)}$$

En realidad ambas formulas son iguales, solo que en el rpimer caso el termino √(N-n)/N se aproxima a uno. Además, en caso de muestras pequeñas (n<30), remplazamos el término $Z_{(\alpha/2)}$ por $t_{(\alpha/2)}$ .

Para estimar la varianza poblacional utilizaremos el estadístico varianza muestreal:

$$s^2 = \frac{\sum( x_i - \overline{x} )^2}{(n - 1)}$$

El intervalo de confianza vendrá dado por:

$$\frac{(n - 1) s^2}{\chi^2_{(\alpha/2)}} < \sigma^2 < \frac{(n - 1) s^2}{\chi^2_{(1-\alpha/2)}}$$

El estimador del total poblacional $\tau = N\mu$ será:

$$\tau = N x$$

El estimador para la proporción poblacional **p** vendrá dado por la proporción muestreal x/n:

$$p = x / n = \frac{x}{n}$$

y su error de estimación por:

$$Z_{\alpha/2}\sqrt{\frac{\frac{x}{n}(1 - \frac{x}{n})}{n-1}\left(\frac{N-n}{N}\right)}$$

Tamaño de la muestra

Para determinar el tamaño de la muestra utilizaremos la siguiente formula para medias :

$$n = [\frac{Z_{(\frac{\alpha}{2})}\sigma}{\Delta}]^2$$

La cual no es mas que la fórmula del error despejada, y en donde n es el tamaño de la muestra, $\sigma$ es la varianza y $\Delta$ el máximo error que estamos dispuestos a aceptar.

Para proporciones utilizaremos:

$$n = \frac{N(p)(1-p)}{(N-1)\Delta - p(1-p)}$$

Lógicamente que estas fórmulas las deberemos aplicar antes de realizar el muestreo, en cuyo caso desconoceremos $\sigma$ y p. Pero estos valores se los pueden obtener de poblaciones similares, muestreos anteriores a dicha población, o un muestreo de prueba. Sin embargo, en la segunda ecuación podemos también remplazar p por 0.5 para obtener un tamaño de muestra conservador.

Muestreo Aleatorio Estratificado

Una muestra aleatoria estratificada es la obtenida mediante la separación de los elementos de la población en grupos que no presenten traslapes, llamados estratos, y la selección posterior de una muestra irrestricta aleatoria simple en cada estrato.

Ya que nuestro objetivo al diseñar un muestreo es maximizar la información obtenida a un costo dado, este tipo de muesteo puede ser mas eficiente que el totalmente aleatorio bajo ciertas condiciones. Básicamente lo que se hace es seleccionar estratos dentro de la población donde suponemos que la información va a ser mas homogénea que en la población en general.

El primer paso en la selección de una muestra aleatoria estratificada es <u>especificar claramente los estratos</u>. Cada unidad muestreal se ubica en el estrato apropiado, y es importante definir correctamente el estrato y que cada unidad muestreal este en uno y solo un estrato. Después que las unidades de muestreo han sido divididas en estratos, seleccionamos una uestra totalmente aleatoria en cada estrato mediante la técnica que ya describimos en la sección anterior. Es importante que las muestras seleccionadas en cada estrato sean independientes. Es decir que las muestras seleccionadas en un estrato no dependan de las seleccionadas en otro.

Definiremos:

- **Número de estratos:** L
- **Numero de unidades muestreales en el estrato i:** Ni
- **Número de unidades muestreales en la población:** N = ΣNi
- **Tamaño de la muestra en el estrato i:** ni
- **Media del estrato i:** μi
- **Media de la población:** μ
- **Variaza del estrato i:** $\sigma^2$i
- **Varianza de la Población:** $\sigma^2$
- **Total del estrato i:** τ i
- **Total Poblacional:** τ

<u>Estimación media, Varianza, error</u>

Sea xi el promedio para la media muestreal del estrato i, ni el tamaño de la muestra del estrato i, μi la media del estrato i y τi el total poblacional para el estrato i, entonces el total de la población es τ = τ1 +τ2+ .. τi+ ... τ L. Ya que tenemos una muestra totalmente dentro de cada estrato, y que xi es un estimador insesgado de μi, y que Ni xi es un estimador insesgado de τ = Niμi, es razonable tener un estimador de τ mediante la suma de los estimadores de τi (Σ Nixi). Asi mismo, ya que el promedio es igual al total polacional dividido para N (τi/N), un estimador insesgado de μ se puede obtener sumando los los estimadores de τi de todos los estratos y luego dividiedolos para N, y llamando a este estimador xst, en donde st indica que se ha utilizado un muestreo aleatorio estratificado:

$$x_{st} = \frac{1}{N}\sum_{i=1}^{L} N_i \bar{x}_i$$

Que como podemos apreciar es bastante parecido a la formula del promedio ponderado.

El estimador de la varianza de xst será:

$$\sigma^2 = \frac{1}{N^2}\sum_{i=1}^{L} N_i^2\left(\frac{N_i - n_i}{N_i}\right)\left(\frac{s_i^2}{n_i}\right)$$

y el límite para el error de estimación E :

$$E = Z_{\alpha/2}\sqrt{\frac{1}{N^2}\sum_{i=1}^{L} N_i^2\left(\frac{N_i - n_i}{N_i}\right)\left(\frac{s_i^2}{n_i}\right)}$$

Para proporciones, el estimador de la proporción poblacional p vendrá dado por:

$$p_{st} = \frac{1}{N}\sum_{i=1}^{L} N_i p_i$$

Y los límites para el error de estimación por:

$$E = Z_{\alpha/2}\sqrt{\frac{1}{N^2}\sum_{i=1}^{L} N_i^2\left(\frac{N_i - n_i}{N_i}\right)\left(\frac{pq}{n_i - 1}\right)}$$

Tamaño de la muestra

El problema del tamaño de la muestra se complica un poco al tratar con muestreos estratificados al azar respecto al de muestreo totalmente aleatorio, ya que no solo vamos a querer conocer el tamaño de la muestra total n, si no queremos saber que porcentaje de la muestra corresponde a cada estrato, osea el tamaño de la muestra ni y/o la proporción wi=ni/n. Además, puede ser que el costo de muestrear cada estrato sea distinto.

Examinemos un método para seleccionar el tamaño de la muestra, a fin de obtener una cantidad fija de información para estimar la media $\mu$. O sea que queremos que el error de estimación sea no mas de E = $Z_{\alpha/2}\ \sqrt{\sigma^2}$. Pero primero debemos de conocer las fracciónes wi = ni/n. Cada manera de asignar esta fracción puede originar una varianza distinta. En términos generales, el mejor método de asignación está influido por:

1. El número total de elementos en cada estrato (Ni)
2. La variabilidad de las observaciones dentro de cada estrato.

3. El costo de obtener una observación de cada estrato.

Ni afecta la cantidad de información en la muestra, ya que mientras mas grande sea, necesitaremos mas observaciones para obtener una cantidad de información.
La variabilidad afecta, porque mientras mayor sea necesitaremos mas observaciones.
Si el costo de obtener una observación varía de un estrato a otro, se tratará de tomar muestras lo mas pequeñas posibles de estratos con costos altos.

No vamos a tomar en cuenta el costo en este curso. Si se desea ampliar este tema se puede consultar a Scheaffer y Mendenhall (1986), los cuales dan una excelente descripción de el. Si suponemos que el costo de obtener la información es el mismo (lo cual puede ocurrir en muchos casos) , entonces la proporción w será:

$$w_i = \frac{N_i \sigma_i}{\sum_{i=1}^{L} N_i \sigma_i}$$

Una vez conocida esta, puede calcular el valor de n como:

$$n = \frac{\left(\sum_{i=1}^{L} N_i \sigma_i\right)^2}{N^2 \left( E / Z_{\alpha/2} \right)^2 + \sum_{i=1}^{L} N_i \sigma_i}$$

Y calcular los valores de ni como:

**ni = nwi**

<u>Regresión Lineal.</u>
Llamamos regresión a un problema en el cual fijamos valores dados de una variable independiente (x), y realizamos observaciones en una variable dependiente (y) de ésta.

El propósito de este estudio es lograr una ecuación para predecir y a partir de x, dentro de un rango específico.

En un análisis de correlación se mide, para cada muestra los valores de x y y; éstos son graficados para encontrar relaciones entre ellos. Además se calculan algunos estadísticos para determinar la fuerza de la relación, aunque un alto valor de correlación no indica necesariamente que x es causado por y, o viceversa.

Por lo tanto, la regresión puede ser usada para experimentos reales, mientras que la correlación se usa para estudios ex post facto, es decir, para analizar datos obtenidos con anterioridad.

Además de una herramienta descriptiva, la regresión puede ser usada como una herramienta comparativa, y como una herramienta predictiva, pudiéndose incluso asignar límites del error para las predicciones. Esto lo veremos más adelante.

Diagrama de dispersión.-

Llamamos diagrama de dispersión a un gráfico en el cual van a estar representados, mediante puntos, los valores de nuestros pares de variables (x,y).

Este diagrama sirve para darnos una idea visual del tipo de relación que existe entre ambas variables, y debe de ser hecho antes de iniciar cualquier cálculo para evitar trabajos innecesarios.

Método de los mínimos cuadrados.-

Llamamos regresión lineal a un experimento donde tratamos de relacionar dos variables x y y, mediante una ecuación de la recta, esto es:

**y = a + bx**

en donde **a** es la intersección de la recta con el eje Y, y **b** es la pendiente de la recta.

Para encontrar los valores de a y b de la recta que más se acerca a nuestros datos experimentales, utilizamos el método de los mínimos cuadrados; es decir, vamos a tomar la recta para la cual los cuadrados de las diferencias entre los puntos experimentales (x,y) y los puntos calculados (x',y') sea mínima.

Las fórmulas para calcular los coeficientes a y b son:

$$b = \frac{\sum xy - \frac{\sum\sum x \sum\sum y}{N}}{\sum\sum x^2 - \frac{(\sum\sum x)^2}{N}}$$

y:

$$a = \frac{\sum\sum y}{N} - b\frac{\sum\sum x}{N}$$

Coeficientes de correlación.-

Llamamos coeficiente de determinación ($r^2$) a la proporción de la variación en la variable y que puede ser atribuida a una regresión lineal con respecto a la variable x. Se lo calcula mediante la fórmula:

$$r^2 = \left( \frac{N\sum xy - (\sum x \sum y)}{\sqrt{[N\sum x^2 - (\sum x)^2][N\sum y^2 - (\sum y)^2]}} \right)^2$$

Su raíz cuadrada positiva (r) se la conoce como coeficiente de correlación de Pearson y es un estimador del parámetro coeficiente de correlación poblacional $\rho$.

Llamamos eta cuadrado ($\eta^2$) a la relación entre la suma de cuadrados entre tratamientos (SCT) y la suma de cuadrados total (SC Total) del ANOVA, y representa a la máxima variación total que puede ser atribuida a cualquier regresión de y con respecto de x; o sea, la máxima correlación que podemos esperar en una curva o modelo matemático que pase por todas las y para cada valor de x.

La regresión la veremos en esta sección como una medida descriptiva de la relación de dos variables, pero también se la utilizará mas adelante como una herramienta comparativa y como un medio de predicción.

Veamos un ejemplo del cálculo de la regresión:

Los siguientes son valores de concentración de cloro (en ppm) a diferentes horas después de la aplicación. Calcule una ecuación que describa el comportamiento de la concentación de cloro respecto al tiempo.

| **horas (x)** | **ppm Cl- (y)** | **$x^2$** | **$y^2$** | **xy** |
|---|---|---|---|---|
| 2 | 1.8 | 4 | 3.24 | 3.6 |
| 4 | 1.5 | 16 | 2.25 | 6.0 |
| 6 | 1.4 | 36 | 1.96 | 8.4 |
| 8 | 1.1 | 64 | 1.21 | 8.8 |
| 10 | 1.1 | 100 | 1.21 | 11.0 |
| 12 | 0.9 | 144 | 0.81 | 10.8 |
| 42 | 7.8 | 364 | 10.68 | 48.6 |

N = 6

Las fórmulas para calcular los coeficientes a y b son:

$$b = \frac{48.6 - \frac{(42)(7.8)}{6}}{364 - \frac{(42)^2}{6}} = -0.0857$$

b = -0.0857

$$a = \frac{7.8}{6} - -0.0857\frac{42}{6} = 1.9$$

a = 1.9

Y la ecuación será:

$y = 1.9 - 0.0857x$

Si queremos calcular el valor de [Cl] a las 3.5 horas, solo reemplazamos x por 3.5:

$y = 1.9 - 0.0857\ (3.5) = 1.60005$ ppm

Calculamos además el valor de $r^2$:

$$r^2 = \left( \frac{n\Sigma xy\ (\Sigma x \Sigma y)}{\sqrt{((n\Sigma x^2\ (\Sigma x)^2)(n\Sigma y^2\ (\Sigma y)^2)}} \right)^2 = \frac{36 * 0.97590}{36.88902} =$$

$r^2 = 0.952380$

<u>Regresiones no lineales .-</u>

Además de la regresión lineal existen otros tipos de relaciones posibles entre las variables x y y. En los fenómenos naturales de crecimiento poblacional es muy común la regresión exponencial de la forma:

$$y = ab^x$$

en donde a es conocido como "índice de Falton", y b es el índice de crecimiento relativo.

Es característico de esta relación que, al graficarse sus pares de datos (x,y) en un papel semilogarítmico, el resultado sea una línea recta.

Podemos, de la misma forma, enderezar los datos numéricamente reemplazando esta ecuación por:

$$\log y = \log a + x \log b$$

O sea, linealizando la curva, después de lo cual tendremos un caso de regresión lineal.

Otro caso de regresión no lineal o curvilínea, ocurre cuando la relación entre ambas variables sigue una curva de grado 2 ó mayor, o sea que sigue una ecuación polinomial de la forma :

$$y = b_0 + b_1 x + b_2 x^2 + \ldots + b_p x^p$$

El ajuste de curvas polinomiales también se utiliza para obtener aproximaciones mediante regresión, cuando nuestro modelo lineal no tiene suficiente fuerza. Pero, no estudiaremos este caso en nuestro curso.

## Bondad de Ajuste

Para usar las diferentes distribuciones teóricas que estudiamos, en nuestros problemas prácticos, debemos primero de asegurarnos que nuestros datos se aproximen a una de estas distribuciones dadas.
Muchas de las pruebas que vamos a usar más adelante (como por ejemplo ciertas pruebas de hipótesis y los análisis de varianza) están basadas en la suposición de que nuestra población está distribuida de tal forma que sigue una distribución normal. Si esta condición no se cumple, entonces no las podremos usar.
Es por esto que estudiaremos a continuación una prueba de bondad de ajuste, la cual nos permitirá decir si nuestros datos observados siguen una distribución teórica dada. Otra aplicación práctica de la bondad de ajuste es para poder describir una población. Con la bondad de ajuste podemos determinar que tan bien se ajusta una población dada a una distribución dada, y describir dicha población de esta forma.

Hablamos de bondad de ajuste cuando tratamos de comparar una distribución de frecuencia observada con los correspondientes valores de una distribución esperada o teórica.

Estudiaremos la prueba Ji-cuadrado para bondad de ajuste, la cual sirve tanto para distribuciones discretas como para distribuciones continuas.

Veamos el procedimiento a seguir con un ejemplo:

Los siguientes son datos de distribución de frecuencias de longitud standard en mm. de alevines de Lebistess sp, cuyo promedio es igual ax =18.55 y su desviación estándar muestreal s=5.55.:

| *Interv. Represen* | | *Frecuencia observada* |
|---|---|---|
| *li* | *ls* | |
| 5.0 | 8.9 | 4 |
| 9.0 | 12.9 | 10 |
| 13.0 | 16.9 | 14 |
| 17.0 | 20.9 | 25 |
| 21.0 | 24.9 | 17 |
| 25.0 | 28.9 | 9 |
| 29.0 | 32.9 | 4 |

Queremos probar si con un 95% de confianza estos datos siguen una distribución Normal con $\mu$=18.55 y $\sigma$=5.55.

Siguiendo el procedimiento:

1.- Ordenamos nuestros datos separándolos en k rangos o clases, cuidando de que en cada rango (i) la frecuencia observada (oi) sea > 4, lo que ya está hecho en la tabla.

2.- Contamos las frecuencias observadas en cada clase (oi), que también está hecho.

3.- Decidimos la distribución a la cual queremos ajustar nuestros datos, expresando la hipótesis nula y su alterna:

Decíamos que queremos ver si nuestros datos se aproximan a una distribución Normal con $\mu$=18.85 y $\sigma$=5.55.

Entonces:

H0 = Hay buen ajuste de nuestros datos a la distribución N(18.55,5.55)
H1 = No Hay buen ajuste nuestros datos a la distribución N(18.55,5.55)

4.- Calculamos la frecuencia teórica esperada ei para cada intervalo i, siendo ésta igual al producto del tamaño muestreal N por la probabilidad de dicho rango obtenida de la tabla correspondiente:

| Interv. Repres. | | oi | P(li≥Z≥ls) | ei |
|---|---|---|---|---|
| 5.0 | 8.9 | 4 | 0.0375 | 3.1 |
| 9.0 | 12.9 | 10 | 0.1071 | 8.9 |
| 13.0 | 16.9 | 14 | 0.2223 | 18.5 |
| 17.0 | 20.9 | 25 | 0.2811 | 23.3 |
| 21.0 | 24.9 | 17 | 0.2163 | 18.0 |
| 25.0 | 28.9 | 9 | 0.1013 | 8.4 |
| 29.0 | 32.9 | 4 | 0.0344 | 2.9 |
| T O T A L | | 83 | 1.0000 | 83.1 |

5.- Calculamos el estadístico $\chi^2$ :

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

El cual sigue una distribución Ji-cuadrada con $\nu$ = k - m - 1 grados de libertad, en donde k es el número de intervalos y m es el número de parámetros estimados. Como estamos tratando de aproximar a una distribución Normal utilizaremos m=2 (media y varianza) y grados de libertad = n = 7 - 2 - 1 = 4.

Para simplificar y visualizar mejor los cálculos, y como ya tenemos el cuadro construido, podemos calcular el valor de c2 para cada intervalo y luego sumarlos todos:

| Interv. Repres. | | oi | P(li≥Z≥ls) | ei | $\chi^2$ |
|---|---|---|---|---|---|
| 5.0 | 8.9 | 4 | 0.0375 | 3.1 | 0.2613 |
| 9.0 | 12.9 | 10 | 0.1071 | 8.9 | 0.1360 |
| 13.0 | 16.9 | 14 | 0.2223 | 18.5 | 1.0946 |
| 17.0 | 20.9 | 25 | 0.2811 | 23.3 | 0.1240 |
| 21.0 | 24.9 | 17 | 0.2163 | 18.0 | 0.0556 |
| 25.0 | 28.9 | 9 | 0.1013 | 8.4 | 0.0429 |
| 29.0 | 32.9 | 4 | 0.0344 | 2.9 | 0.4172 |
| T O T A L | | 83 | 1.0000 | 83.1 | 2.1315 |

6.- Si el valor de $\chi^2$ calculado es menor que el correspondiente valor de $\chi^2$(0.05) para 4 grados de libertad (9.4880) de la tabla, concluimos que existe un buen ajuste; de lo contrario, no.

W={2.1315 > 9.4880)

Como esto es falso, no rechazamos la hipótesis, pero la aceptamos y concluimos que:
Si hay un buen ajuste de nuestros datos a la distribución Normal con $\mu$=18.55 y $\sigma$=5.55.

Otro uso práctico que podemos hacer de el método de bondad de ajuste es determinar distribuciones de tamaño para distintos tamaños promedio. De esta forma podemos aproximar nuestras distribuciones y obtener una distribución en base un promedio y una varianza, lo cual facilita la realización de modelos para determinar momento óptimo de cosecha, ya que podemos calcular así el valor esperado de venta a determinado tamaño.

Hay otros métodos de bondad de ajuste. Un método fácil y práctico para determinar si una distribución es normal es la prueba de la Kurtosis y la skewness (sesgo).

Para esta prueba calculamos el estadístico Skewness:

$$b_1 = \frac{\sqrt{n} x \sum( x_i - \bar{x} )^3}{( \sum( x_i - \bar{x} )^2 )^{3/2}}$$

El cual es sensible al sesgo o desviaciones en simetría de la normal. Valores positivos indican una cola derecha mas larga y valores negativos una cola izquierda mas larga. Un valor de 0 indica una simetría perfecta. Los valores críticos para diferentes valores de n se encuentran en este gráfico. Valores absolutos calculados de b1 mayores que los del gráfico corresponden a desviaciones significativas al correspondiente nivel de confianza en cuanto a sesgo.

Y el estadístico Kurtosis:

$$b_2 = \frac{n x \sum( x_i - \bar{x} )^4}{( \sum( x_i - \bar{x} )^2 )^2}$$

El cual es sensible a desviaciones en la distribución y el pico de la curva. Valores > a 3 indican distribuciones leptokurtoticas con un pico mas alto y desviaciones menores que la normal. Valores < a 3 pero > a 1 + b1 indican distribuciones platikurtóticas, con un pico mas bajo y desviaciones mayores que la normal.
Valores calculados fuera de los límites de este gráfico indican desviaciones significativas en la kurtosis al nivel de significancia correspondiente.

Como ejemplo realizamos esta prueba para el ejercicio anterior.

b1 = 0.448

Su valor absoluto es mayor que 0.28, el valor para n=200 en el gráfico para a = 0.05, por lo cual decimos que si hay desviaciones significativas en sesgo con respecto a la normal.

b2 = 2.115

Su valor es menor que 2.42, el límite inferior para n = 200 y a = 0.05 en el gráfico correspondiente, por lo cual decimos que si hay desviaciones significativas en kurtosis respecto a la normal.

Entonces concluimos que estos datos no siguen una distribución normal. Que es el mismo resultado que obtuvimos de la prueba Ji-cuadrado, aunque no necesariamente deberán coincidir ambos resultados.

**Estadística Comparativa.-**

La estadística comparativa es aquella parte de la estadística que se propone comparar dos o mas poblaciones. Existen algunas herramientas para hacer comparaciones. Las mas conocidas son las pruebas de hipótesis y el análisis de varianza.

Empezaremos dando una breve explicación de lo que es una prueba de hipótesis y como funcionan para seguir con la manera de realizar cada prueba y cuando usarlas.

Llamamos hipótesis estadística a una asumción sobre una población que está siendo muestreada.

Un test de hipótesis es simplemente una regla mediante la cual esta hipótesis se acepta o se rechaza.
Esta regla está basada generalmente en un estadístico muestreal llamado estadístico de prueba, ya que se lo usa para probar la hipótesis.

La región crítica de un estadístico de prueba consiste en todos los valores del estadístico donde se hace la decisión de rechazar H0.

Debido a que las pruebas de hipótesis están basadas en estadísticos calculados a partir de *n* observaciones, la decisión tomada está sujeta a posibles errores.
Si rechazamos una hipótesis nula verdadera, estamos cometiendo un error de tipo I. La probabilidad de cometer un error del tipo I se llama $\alpha$.
Si aceptamos una hipótesis nula falsa, estaremos cometiendo un error del tipo II, y la probabilidad de cometerlo se la denomina $\beta$.

El siguiente cuadro denota esto:

| HECHO | D E C I S I O N | |
|---|---|---|
| | Se acepta H0 | Se rechaza H0 |
| H0 verdadera | Decisión correcta | **Error de tipo I** |
| H0 falsa | **Error de tipo II** | Decisión correcta |

Uno de los objetivos de las pruebas de hipótesis es diseñar tests en donde $\alpha$ y $\beta$ sean pequeños, pero la mayor ventaja que nos dan las pruebas de hipótesis es precisamente esto: Poder medir $\alpha$ y $\beta$, de tal forma que nosotros podamos medir la incertidumbre, remplazando palabras vagas como "pudo" "tal vez" posiblemente" que nosotros ponemos al "ojímetro" por un número que denota cuanto es la posibilidad de equivocarnos.

Para probar una hipótesis, generalmente la expresamos en su forma nula (**H0**), y formulamos una hipótesis alterna (**H1**) que aceptaremos al rechazar la nula.
Expresar una hipótesis de su forma nula significa poner de tal forma que no haya diferencias. Además debe de ser una hipótesis simple, por ejemplo:
"No hay diferencia entre las medias"
"La media de la población es igual a 10"
"Los caracteres son independiente"
"Las varianzas son homogéneas(iguales)"
etc.
Ambas hipótesis deben ser distintas y mutuamente excluyentes.

Un ejemplo de una prueba de hipótesis resultaría si nosotros quisiéramos saber si dos poblaciones, una con peso promedio de muestreo 14.0 g y otra con 15.0 g tienen igual media de peso. Obviamente nosotros diremos "14.0 no es lo mismo que 15.0", pero entonces: cuando son iguales? cuando la primera piscina pesa 14.2? 14.5g ? 14.8? 14.9? 14.99? porque no 14.85? o 14.849?.
Esto no lo podemos saber al ojo o por "feeling", tenemos que usar un método estadístico que nos diga si son o no diferente con un $(1-\alpha)$x100% de confianza. Y esto es precisamente lo que hacen las pruebas de hipótesis.
Básicamente para medias, el punto donde cambia de igual a diferente va estar dado de acuerdo a la varianza de las poblaciones y al tamaño de la muestra. Lógicamente, en una población donde la varianza es grande y hemos tomado una muestra pequeña, la probabilidad de que este promedio represente a la media va a ser menor que en una población donde la varianza es pequeña y el tamaño de la muestra grande.

El % de confianza $(1-\alpha) \times 100$ % no es mas que decir que tenemos este porcentaje de confianza de no estar cometiendo un error del tipo I, o que si repitiéramos infinitamente el experimento en las mismas condiciones, al menos este porcentaje de veces nos daría el mismo resultado.
Y el área crítica(W) es el área de la curva en donde H0 va a ser rechazada en $\alpha$x100% de las veces.

Antes de ver los pasos a seguir para resolver una prueba de hipótesis veamos como funcionan en teoría con un ejemplo:

Supongamos que deseemos saber si la media de concentración de cierta sustancia en un alimento no excede 20 mg/g (ya que a concentraciones mayores produce efectos indeseables).
Bueno, para averiguarlo empezamos a tomar muestras de este alimento y a hacer análisis para medir la sustancia, pero como era de esperarnos va a haber cierta variación en los resultados, (si no la hubiera no necesitaríamos la estadística) puesto que provienen de una población que tiene cierta varianza.
Supongamos que nosotros conocemos que la varianza de esta población es de $\sigma^2 = 5.95$ (lo conocemos por experiencias pasadas, por este muestreo, o por revelación divina, pero en este ejemplo la conocemos).
Bueno, seguimos muestreando hasta obtener n = 36 datos (no importa ahora como fue que decidimos este tamaño de muestra).
y resulta que nuestro promedio nos da$x$ = 20.75 mg/g.

La pregunta del millón es: ¿Es este promedio lo suficientemente alto para ser considerado "mayor"? o en otras palabras :¿Cual es la probabilidad de que en un muestreo de 36 observaciones de una población con media $\mu$=20 y varianza $\sigma^2$=5.95 obtengamos un promedio de $x$ =20.75 o más?. Y si Uds. se acuerdan, cuando estudiamos las distribuciones derivadas, ya resolvimos este problema.

Dijimos que la media de la población de promedios es igual a la media de la población de donde fue tomada $\mu x=\mu$ y la varianza es igual a un enésimo de la varianza de la población principal $\sigma^2=\sigma^2/n$. Entonces, los parámetros de nuestra distribución derivada de medias de esta población, en caso de que $\mu$=20 mg/g serán:

$\mu x$ = 20 mg/g

$\sigma^2 = 5.95/36 = 0.165$
$\sigma = 0.4$
De donde solo queda determinar:

$$P(x \geq 20.75)$$

Tipificando este valor tendremos:

$$Z = \frac{x - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{20.75 - 20}{0.4} = 1.875$$

y, buscando en la tabla tenemos:

$$P(Z \geq 1.875) = 0.5 - 0.4696 = \underline{0.0304}$$

Lo que significa que si hemos muestreado de una población con $\mu$=20 y $\sigma^2$=5.95, la probabilidad de obtener un promedio de 20.75 o mas es del 3%, lo cual es considerado "poco probable" (si trabajamos con a=0.05), o en otras palabras, la probabilidad de <u>erróneamente</u> decir que "la media de este alimento balanceado es mayor que 20 mg/g" (rechazar H0) es 3% ($\alpha$=0.03)(probabilidad de cometer un error de tipo I). Este sería el riesgo que correríamos si fuéramos los productores del alimento y rechazáramos vender un producto bueno por considerarlo malo cuando en verdad esta bueno.

En la práctica no se hacen todos estos cálculos para una prueba de hipótesis, si no que se usan pruebas “rutinarias” que ya están establecidas según lo que queramos comparar. Las ”recetas” o fórmulas que se usan para cada tipo de comparación las podemos encontrar en cualquier libro de estadística.

El valor de $\alpha$ al cual decidimos nosotros aceptar o rechazar H0 va a depender únicamente de nosotros, que tanto deseamos estar seguros de no equivocarnos. Si el resultado de equivocarme va a resultar en que Yo me muera me inclino por valores de $\alpha$ bajos (0.00000000001), si el resultado de equivocarme va a resultar en que a mi perro le salgan canas me iría por un valor mas alto (x ej. $\alpha$=0.1). En general se usan valores entre 0.1 y 0.01, siendo el mas común el de 0.05.

Los pasos básicos para efectuar la mayoría de las pruebas de hipótesis son los siguientes:

1.- Expresar claramente la hipótesis nula (H0) y su alterna (H1), que están dadas en los libros para algunas de las pruebas más comunes.

2.- Especificar el nivel de significancia $\alpha$ y el tamaño de la muestra (n), $\alpha$ se lo asigna generalmente en 0.05 ó 0.01, de acuerdo a la precisión que deseemos, y n se verá más adelante como se la asigna.

3.- Escoger un estadístico para probar H0, tomando en cuenta las asumciones y restricciones que involucran usar este estadístico, esto también está dado en los libros para las distintas pruebas existentes.

4.- Determinar la distribución muestreal de este estadístico cuando H0 es verdadera, está dado en los libros la distribución muestreal del estadístico para cada prueba.

5.- Designar la región crítica de la prueba, en la cual H0 va a ser rechazada en 100 x $\alpha$% de las muestras cuando H0 es verdadera, es necesario calcularla dependiendo de los datos que se estén probando, pero en los libros se indica en general entre que valores está definida.

6.- Escoger una (dos) muestra(s) aleatoria(s)de tamaño n, dependiendo si estamos probando una o dos poblaciones, este proceso es simplemente mecánico y debe de realizarse aleatoriamente, Midiendo la variable que nos interese.

7.- Calcular el estadístico de prueba, lo cual es simplemente remplazar los datos obtenidos en la fórmula dada en el libro.

8.- Por último comparar el estadístico calculado con el teórico y decidir en base al resultado y guiándonos por la zona crítica o de rechazo si:
a) aceptamos H0.
b) rechazamos H0 (y aceptamos H1).
c) no tomamos ninguna decisión (si pensamos que los datos no son concluyentes).

Empezaremos con las pruebas unimuestrales en las que tratamos de probar si un parámetro calculado a partir de un estadístico es igual o no a un valor predeterminado o a un parámetro poblacional conocido.

Estudiaremos cuatro pruebas en este capítulo: una media con varianza conocida, una media con varianza desconocida, una varianza y una proporción.

Una media con varianza conocida.-

Utilizamos esta prueba para comparar una media poblacional conocida ($\mu$) con la calculada a partir de el promedio de una muestra cuya varianza conocemos($\mu_0$); o, que en su defecto, su tamaño sea n $\geq$ 30, por lo que supondremos que la varianza poblacional sea igual a la muestreal.

Las hipótesis a probar son:

H0 = $\mu_0 = \mu$
H1 = $\mu_0 \neq \mu$

El estadístico de prueba usado es:

$$Z = \frac{\overline{x} - \mu}{\sigma / \sqrt{n}}$$

donde Z sigue una distribución normal tipificada (0,1).

La región crítica o de rechazo (W) viene dada por:

$|Z| \geq Z_{(\alpha/2)}$

Se pueden probar también estas hipótesis alternas con sus respectivas regiones de rechazo (W):

$H0 = \mu_0 = \mu$

$H1 = \mu_0 < \mu$ $\qquad W = \{Z \leq -Z_{(a)}\}$

ó,

$H0 = \mu_0 = \mu$

$H1 = \mu_0 > \mu$ $\qquad W = \{Z \geq Z_{(a)}\}$

Por ejemplo: Supongamos que estamos empacando fundas de 40 Kg. de alimento balanceado. Queremos saber si se están pesando bien los sacos. Aquí no podemos empacar más de 40 Kg. por que resultaría en pérdida, ni podemos empacar menos de 40 Kg. por que los clientes se darían cuenta y se irían a la competencia.
Queremos determinar si se está pesando bien.

Aquí las hipótesis a probar serían

$H0 = \mu_0 = 40$

$H1 = \mu_0 \neq 40$

Decidimos trabajar con a=0.05 y tomar una muestra de n = 45 sacos que son equivalentes a una parada.

Sabemos que el estadístico a usar es Z, el cual sigue una distribución normal tipificada, y cuya región crítica es:

$W = \{|Z| \geq 1.96\}$

Tomamos la muestra de n= 45 y obtenemos los siguientes estadísticos:

x = 39.8 Kg.

$s^2 = 4.2$

Como n > 30, reemplazamos $\sigma^2$ por $s^2$, por lo que $\sigma$=2.05.

calculamos Z = (39.7 - 40)/(2.05/√45)= -0.3/0.306= -0.98

como

$$W = \{ |-0.98| \geq 1.96\}$$

es falso, no rechazamos la hipótesis nula, sino que la aceptamos y concluimos que el peso medio de los sacos no es mayor ni menor que 40 sino igual.

<u>Una media con varianza desconocida.-</u>

Utilizamos esta prueba para comparar una media poblacional conocida ($\mu$) con la calculada a partir de el promedio de una muestra ($\mu_0$) cuya varianza no conocemos, y cuyo tamaño sea < 30. En este caso trabajaremos con la varianza muestreal en vez de la poblacional, pero usaremos el estadístico de prueba "t".

Las hipótesis a probar son:

H0 = m0 = m
H1 = m0 ¹ m

El estadístico de prueba usado es:

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$

donde t sigue una distribución "t" de Student con n-1 grados de libertad.

La región crítica o de rechazo (W) viene dada por:

$$|t| \geq t_{(a/2)}$$

Se puede probar también estas hipótesis alternas con sus respectivas regiones de rechazo (W):

$H0 = \mu_0 = \mu$
$H1 = \mu_0 < \mu$ $\qquad W = \{t < -t_{(\alpha)}\}$

ó,
$Ho = \mu_0 = \mu$
$H1 = \mu_0 > \mu$ $\qquad W = \{t \geq t_{\alpha}\}$

Supongamos que deseamos medir la concentración de cierto químico en un producto. Suponiendo que el límite de concentración de dicho químico sea de 100 ppm. Que decisión tomamos?

Aquí hay otra consideración que hacer, ya que suponiendo que los análisis son costosos, y además el producto se destruye en cada análisis, no podemos tomar una muestra demasiado grande.

Usaremos las hipótesis:

Ho = $\mu_0$ = 100
H1 = $\mu_0$ > 100

Trabajaremos con un nivel de significancia de $\alpha$=0.05 y una muestra de tamaño n=5.

Como desconocemos la varianza poblacional, trabajaremos con el estadístico t, el cual sigue una distribución "t" de Student con $\nu$=5-1=4 grados de libertad.

La región crítica viene dada por:

$$W = \{t \geq 2.13\}$$

Tomamos la muestra de n= 5 y obtenemos los siguientes estadísticos:

x = 105.6 ppm
s2= 17.3

Utilizamos $s^2$ en vez de $\sigma^2$ por lo que s=4.16

calculamos t = (105.6 - 100)/(4.16/√5)= 5.6/1.86= 3.01

como

$$W = \{3.01 \geq 2.13\}$$

es verdadero, rechazamos la hipótesis nula, y aceptamos la alterna, concluyendo que la concentración de químico en el producto es mayor que 100 ppm.

<u>Una varianza.-</u>

Utilizamos esta prueba para comparar una varianza poblacional conocida ($\sigma^2$) con una calculada a partir de la varianza muestreal de una población muestreada ($\sigma^2_0$), suponiendo con cierto grado de confianza que esté normalmente distribuida. En este caso trabajamos con el estadístico de prueba $\chi^2$.

Las hipótesis a probar son:
Ho = $\sigma^2_0 = \sigma^2$
H1 = $\sigma^2_0 \neq \sigma^2$

El estadístico de prueba usado es:

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

donde $\chi^2$ sigue una distribución Ji - cuadrado con $\nu$ = n-1 grados de libertad.

La región crítica o de rechazo (W) viene dada por:

$\chi^2_{(\alpha/2)} \leq \chi^2 \leq \chi^2_{(1-\alpha/2)}$

Se puede probar también estas hipótesis alternas con sus respectivas regiones de rechazo (W):

$Ho = \sigma^2_0 = \sigma^2$

$H1 = \sigma^2_0 < \sigma^2$ $\quad W = \{\chi^2 \geq \chi^2_{(1-\alpha)}\}$

ó,

$Ho = \sigma^2_0 = \sigma^2$

$H1 = \sigma^2_0 > \sigma^2$ $\quad W = \{\chi^2 \leq \chi^2_{\alpha}\}$

La varianza es una indicación de la calidad de los procesos. Así de un proceso bueno resulta un producto que tiene poca varianza, y de uno malo otro con gran varianza.

No encontré un ejemplo de esta prueba para Acuicultura, por lo que les doy uno de una planta mecánica.

El proceso de bruñido que se utiliza para esmerilar ciertos discos de silicio al grueso apropiado es aceptable solo si $\sigma$ es a lo sumo 0.5. Empleando un nivel de significancia de $\alpha$=0.05, determinar si mediante una muestra de n=15 con s=0.64, podemos decir que el proceso de bruñido es incorrecto.

Aquí, las hipótesis a usar son:

$Ho = \sigma_0 = 0.5$

$H1 = \sigma_0 > 0.5$

El estadístico de prueba usado es $\chi^2$, el cual sigue una distribución Ji-cuadrado con n=15-1=14 grados de libertad.

La región crítica o de rechazo viene dada por:

$W = \{\chi^2 \geq 23.7\}$

Calculamos $\chi^2$= (15-1)0.642/0.52= 5.734/0.25= 22.94

Y, como

$W = \{22.94 \geq 23.7\}$

Es falso, no rechazamos H0, si no que la aceptamos, y decimos que no hay evidencia de que el proceso de bruñido es inapropiado.

<u>Una proporción.-</u>

Se usa para comparar una proporción poblacional conocida p con una calculada a partir de un estadístico $p_0$, donde el producto np debe ser mayor o igual que cuatro.

Aquí las hipótesis a probar son:

Ho: $p_0 = p$
H1: $p_0 \neq p$

Y el estadístico de prueba usado es:

$$Z = \frac{p_o - p}{\sqrt{\frac{pq}{n}}}$$

el cual sigue una distribución normal tipificada (0,1), p es la proporción conocida; $q = 1 - p$ y $p_o$ es la proporción calculada a partir de la muestra.

El área crítica o de rechazo (W) viene dada por:

$|Z| > Z_{(\alpha/2)}$

Se pueden probar también estas hipótesis alternas con sus respectivas regiones de rechazo (W):

Ho = $p_0 = p$
H1 = $p_0 < p$ $\quad W = \{Z < -Z_{(\alpha)}\}$
ó,

Ho = $p_0 = p$
H1 = $p_0 > p$ $\quad W = \{Z > Z_{(\alpha)}\}$

Supongamos que deseamos saber si el porcentaje de P. vannamei en cierto lote de larva silvestre es de al menos 60%. Para esto tomamos n=231 larvas, de las cuales 132 son vannamei, lo que nos da una proporción muestreal x/n=0.57

Las hipótesis a probar son:

$Ho = p_0 = 0.6$
$H1 = p_0 < 0.6$

Comprando las larvas si no rechazamos la hipótesis nula.

Calculamos $Z= (.57-.6)/\sqrt{(.6x.4/231)}= -0.03/0.032= -0.938$

Y, como:

$W = \{-0.938 < -1.645\}$

Es falso, aceptamos la hipótesis nula, y compramos la larva.

Dos poblaciones independientes.-

Llamadas también pruebas bimuestreales, son usadas cuando queremos comparar dos estadísticos poblacionales calculados a partir de muestras de esas poblaciones.

En este capítulo estudiaremos cinco casos: dos varianzas independientes, dos medias independientes con varianzas conocidas, dos medias independientes con varianzas desconocidas e iguales, dos medias independientes con varianzas desconocidas y desiguales y dos proporciones.

Dos Varianzas.-

Utilizamos esta prueba para comparar dos varianzas poblacionales ($\sigma^2_1$ y $\sigma^2_2$), calculadas a partir de las varianzas muestreales ($s^2_1$ y $s^2_2$) de poblaciones muestreadas, suponiendo con cierto grado de confianza que estén normalmente distribuidas. Consideramos $s^2_1$ a la mayor de las dos.

Las hipótesis a probar son:

$Ho = \sigma^2_1 = \sigma^2_2$
$H1 = \sigma^2_1 \neq \sigma^2_2$

El estadístico de prueba usado es:

$$F = \frac{s^2_1}{s^2_2}$$

donde F sigue una distribución F de Fisher - Schnedecor con $\nu_1=n_1-1$ y $\nu_2=n_2-1$ grados de libertad.
La región crítica o de rechazo (W) viene dada por:

$F > F_{(\alpha/2)}$

Se puede probar también esta hipótesis alterna con su respectiva región de rechazo (W):
$Ho = \sigma^2{}_1 = \sigma^2{}_2$
$H1 = \sigma^2{}_1 > \sigma^2{}_2$ $\qquad$ $W = \{F \geq F_{(\alpha)}\}$
Esta prueba es muy usada, ya que es indispensable realizarla antes de la prueba "t" bimuestreal.

Por ejemplo, los siguientes datos corresponden a los pesos promedios finales de 8 piscinas divididas en 2 tratamientos.
Verifique si las varianzas de ambos tratamientos son iguales para $\alpha = 0.1$

| | **Trat # a** | **Trat # b** |
|---|---|---|
| | 12.3 | 13.8 |
| | 14.3 | 11.9 |
| | 13.4 | 15.5 |
| | 15.0 | 15.0 |
| **x** | 13.75 | 14.05 |
| **s** | 1.1676 | 1.6010 |

Las hipótesis a probar son:

$Ho = \sigma^2{}_1 = \sigma^2{}_2$
$H1 = \sigma^2{}_1 \neq \sigma^2{}_2$

El estadístico de prueba usado es F, el cual sigue una distribución F de Fisher con $\nu_1$=4-1=3 y $\nu_2$=4-1=3 grados de libertad.

La región de rechazo viene dada por:

$W=\{F > 9.28\}$

Calculamos el estadístico F= 1.6012/1.16762= 2.56/1.36= 1.88

Y, como

$W=\{1.88 > 9.28\}$

Es falso, aceptamos H0, y concluimos que no hay diferencias significativas.

<u>Medias, varianzas conocidas.-</u>

Utilizamos esta prueba para comparar dos medias poblacionales calculadas a partir del promedio de dos muestras cuyas varianzas conocemos, o en su defecto, cuyo tamaño individual sea ≥ 30, en donde supondremos que la varianza poblacional sea igual a la muestreal.

Las hipótesis a probar son:

$Ho = \mu_1 = \mu_2$
$H1 = \mu_1 \neq \mu_2$

El estadístico de prueba usado es:

$$Z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{\sigma^2_1}{n_1} + \frac{\sigma^2_2}{n_2}}}$$

donde Z sigue una distribución normal tipificada N(0,1).

La región crítica o de rechazo (W) viene dada por:

$|Z| \geq Z_{(\alpha/2)}$

Se puede probar también esta hipótesis alterna con su respectiva región de rechazo (W):

$Ho = \mu_1 = \mu_2$
$H1 = \mu_1 > \mu_2$ W = $\{Z\ \alpha\ Z_{(\alpha)}\}$

Supongamos que queremos ver si la media de peso de 2 piscinas es igual, para esto sacamos una muestra en cada piscina.
Los resultados de dichos muestreos son los siguientes:

$x_1$= 14.2 gr.
$s^2{}_1$= 3.6

$x_1$= 15.1
$s^2{}_2$= 5.7

Las hipótesis a probar son:

$Ho = \mu_1 = \mu_2$
$H1 = \mu_1 \neq \mu_2$

El estadístico de prueba usado es Z el cual sigue una distribución normal tipificada N(0,1).

La región crítica o de rechazo (W) viene dada por:

$|Z| \geq 1.96$

Calculando Z, remplazando $\sigma^2$ por $s^2$ tenemos:

$$Z = \frac{14.2 - 15.1}{\sqrt{\frac{3.6^2}{62} + \frac{5.7^2}{48}}} = \frac{-0.9}{0.94} = -0.957$$

Y, como :

$|-0.957| \geq 1.96$

Es falso, aceptamos H0, y concluimos que no existen diferencias significativas.

<u>Medias, Varianzas desconocidas e iguales.-</u>

Utilizamos esta prueba para comparar dos medias poblacionales calculadas a partir del promedio de dos muestras cuyas varianzas no conocemos, y cuyos tamaños sean < 30, siempre y cuando hayamos demostrado con anterioridad, mediante una prueba F, que las varianzas poblacionales de ambos son iguales.

En este caso trabajaremos con la varianza muestreal en vez de la poblacional, pero usaremos el estadístico de prueba "t".

Las hipótesis a probar son:

Ho = $\mu_1 = \mu_2$
H1 = $\mu_1 \neq \mu_2$

El estadístico de prueba usado es:

$$t = \frac{\bar{x}1 - \bar{x}2}{\sqrt{\frac{(n_1 - 1)s^2{}_1 + (n_2 - 1)s^2{}_2}{n_1 + n_2 - 2} x(\frac{1}{n_1} + \frac{1}{n_2})}}$$

donde t sigue una distribución "t" de Student con n-1 grados de libertad.

La región crítica o de rechazo (W) viene dada por:

$|t| \geq t_{(\alpha/2)}$

Se puede probar también esta hipótesis alterna con su respectiva región de rechazo (W):

Ho = $\mu_1 = \mu_2$
H1 = $\mu_1 > \mu_2$ W = $\{t \geq t_{(\alpha)}\}$

Supongamos que queremos ver si podemos cambiar el horario de alimentación en una camaronera. Al presente estamos alimentando en la tarde por que en pruebas que se hicieron hace algunos años se vio que el crecimiento era mejor así que de mañana.
Que pruebas realizaría?

Si queremos cambiar la alimentación a la noche, la haremos solo si el crecimiento (no tomamos en cuenta en este ejemplo la conversión, pero en un caso real habría que tomarla) es mejor, ya que alimentar de noche conlleva cambios en la logística y aumento de costos.

Esto equivale a las siguientes hipótesis:

Ho = $\mu_1 = \mu_2$
H1 = $\mu_1 > \mu_2$ $\qquad W = \{t \geq t_{(\alpha)}\}$

En donde $\mu_1$ es la media de crecimiento alimentando de noche y $\mu_2$ la media de crecimiento alimentando de tarde.
Haciendo el cambio si rechazamos la hipótesis nula y aceptamos la alterna.
Si queremos alimentar de mañana para reducir personal, ya que al momento tenemos tres parejas de alimentadores para alimentar de 2 a 5 de la tarde, pero pensamos que si alimentamos a lo largo de todo el día lo podemos reducir a dos parejas. Entonces haremos el cambio siempre y cuando el crecimiento (y/o la conversión) no sea menor (y/o mayor) de la que tenemos alimentando de tarde.

Esto equivale a las siguientes hipótesis:

H0 = $\mu_1 = \mu_2$
H1 = $\mu_1 > \mu_2$ $\qquad W = \{t \geq t_{(\alpha)}\}$

En donde $\mu_2$ es la media de crecimiento alimentando de mañana y $\mu_1$ la media de crecimiento alimentando de tarde.
Haciendo el cambio si no rechazamos la hipótesis nula.

El estadístico a usar es el estadístico t, el cual sigue una distribución "t" de Student con $\nu = n_1 + n_2 - 2$ grados de libertad.

Supongamos que los datos del ejercicio que realizamos para igualdad de varianzas corresponden así: El tratamiento a (2) es la alimentación por la mañana y el b (1) la alimentación por la tarde.
Veamos si hay diferencias en crecimiento debido al horario de alimentación ($\alpha$=0.05).
Entonces tenemos:

$x_2$= 13.75 gr.
$s_2$= 1.1676

$x_1$= 14.05 gr.
$s_1$= 1.601

Entonces el estadístico a calcular es "t". El cual sigue una distribución "t" de Student con $\nu$=4+4-2=6 grados de libertad. Siempre y cuando las varianzas de ambas poblaciones sean iguales. Como esto ya lo probamos, entonces seguimos adelante.

La región de rechazo viene dada por:

W={t > 1.94}

El valor del estadístico t viene dado por:

$$t = \frac{14.05 - 13.75}{\sqrt{\frac{(4-1)1.601^2 + (4-1)1.1676^2}{4+4-2} x (\frac{1}{4} + \frac{1}{4})}} = \frac{0.3}{0.9816} = 0.305$$

Y, como:

W={0.305 > 1.94}

Es falso, aceptamos la hipótesis nula y concluimos que no hay diferencias significativas para $\alpha$=0.05.

<u>Medias, varianzas desconocidas y desiguales.-</u>

En el caso de que se haya demostrado la no igualdad de varianzas entre las poblaciones mediante una prueba F, no sería necesario realizar un test de medias, ya que las poblaciones son tan heterogéneas respecto a sus varianzas.

Si, a pesar de esto, se desea realizar una prueba de medias, (problema de Behrens - Fisher), se puede realizar de varias maneras, siendo una de ellas la prueba de Smith - Satterthwaite, usando el estadístico:

$$t' = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s^2_1}{n_1} + \frac{s^2_2}{n_2}}}$$

el cual sigue una distribución "t" de Student con grados de libertad igual a:

$$\nu = \frac{\left(\frac{s^2_1}{n_1} + \frac{s^2_2}{n_2}\right)^2}{\frac{\left(\frac{s^2_1}{n_1}\right)^2}{n_1 + 1} + \frac{\left(\frac{s^2_2}{n_2}\right)^2}{n_2 + 1} - 2}$$

Las hipótesis y áreas críticas son las mismas usadas en la prueba "t" bimuestreal.

No vamos a ver un ejemplo de esta prueba, pero en el caso de que se lo quiera hacer se lo puede hacer fácilmente con un programa estadístico, como las herramientas de Excell.

<u>Dos proporciones independientes.-</u>

En este caso se tienen dos proporciones calculadas a partir de muestras. En este caso las hipótesis a probar son:

Ho = p1 = p2
H1 = p1 ≠ p2

El estadístico de prueba usado es:

$$Z = \frac{p_1 - p_2}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

donde Z sigue una distribución normal tipificada (0,1), y p es la proporción de ambas muestras juntas:

$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

La región crítica o de rechazo (W) viene dada por:

$|Z| \geq Z_{(\alpha/2)}$

Se puede probar también esta hipótesis alterna con su respectiva región de rechazo (W):

Ho = p1 = p2
H1 = p1 > p2 W = $\{Z \geq Z_{(\alpha)}\}$

Por ejemplo queremos ver si nauplios de dos distintos orígenes tienen igual supervivencia. Para esto sembramos un tanque con 2'011,000 nauplios de origen 1 y otro con 2,102,000 Nauplios de origen 2. Cultivamos ambos tanques bajo idénticas condiciones y determinamos la proporción de supervivencia de cada uno. Así, con resultados de cosecha de 1'693,315 y 1'794,825 Pls respectivamente, obtenemos:

x/n1= 1'693,315/2'011,000 = 0.8420

x/n2= 1'794,825/2'102,000 = 0.8539

Las hipótesis a probar son:

Ho = p1 = p2
H1 = p1 ≠ p2

Primero calculamos la proporción de las dos muestras juntas(p):

$$p = \frac{2'011,000x0.8420 + 2'102,000x0.8539}{2'011,000 + 2'102,000} = 0.8481$$

Y luego el estadístico Z:

$$Z = \frac{0.8420 - 0.8539}{\sqrt{0.8481x0.1519(\frac{1}{2'011,000} + \frac{1}{2'102,000})}} =\_33.61$$

donde Z sigue una distribución normal tipificada (0,1)

La región crítica o de rechazo (W) viene dada por:

$|Z| \geq 1.96$

Y, como:

$|-33.61| \geq 1.96$

es verdad, rechazamos H0 y aceptamos H1, esto es concluimos que si hay diferencias significativas entre la proporción de supervivencia de ambos orígenes de nauplios.

<u>Pruebas en muestras dependientes.-</u>

Esta prueba se aplica a muestras de poblaciones dependientes la una de la otra, como en el caso de pruebas en una misma población antes y después de un tratamiento.

La lógica de la prueba se reduce a determinar las diferencias en los individuos de la muestra, y a probar si la media de estas diferencias son iguales a cero.

Las hipótesis a probar son:

Ho: $\mu\Delta = 0$
H1: $\mu\Delta \neq 0$

El estadístico de prueba a usar es:

$$t = \frac{\bar{d}}{s_d / \sqrt{n}}$$

donde t sigue una distribución "t" de Student con n = n-1 grados de libertad es el promedio de las diferencias entre ambas muestras; sd es la desviación estándar muestreal de las diferencias y n es el número de diferencias.

La región de rechazo (W) viene dada por:

$|t| \geq t_{(\alpha/2)}$

Un ejemplo de esto es medir el peso de una muestra de la población antes y después de cierto tratamiento, y comparar con el resultado después del tratamiento.
Por ejemplo, se desea determinar la eficiencia de cierto químico como bactericida en tanques de larvas, para lo cual se realizan los siguientes contajes de bacterias en el agua de 12 tanques antes y después de su aplicación:

| **ANTES** | **DESPUES** |
|---|---|
| 15.200 | 16.300 |
| 10.100 | 8.300 |
| 6.400 | 28.100 |
| 4.300 | 6.800 |
| 9.300 | 11.400 |
| 12.800 | 34.000 |
| 8.400 | 9.300 |
| 6.500 | 4.200 |
| 3.100 | 12.300 |
| 2.400 | 4.500 |
| 9.400 | 4.200 |
| 6.300 | 11.100 |

Determine si hay diferencias significativas (a = 0.01) debidas al tratamiento.

Las hipótesis a probar son:

Ho: $\mu\Delta = 0$
H1: $\mu\Delta \neq 0$

Las diferencias son:
6800, 7000, 3600, 4100, 3300, 15800, 1900, 2300, -100, 7200, 6500, 22900,

$\Delta$ = 6775
$s_d$ = 6481.74
n = 12

El estadístico de prueba a usar es:

$$t = \frac{6775}{6481.74 / \sqrt{12}} = 3.62$$

donde t sigue una distribución "t" de Student con n=12-1=11 grados de libertad; como:

$W=\{ |3.62| \geq 2.20\}$

Es verdadero, rechazamos H0 y aceptamos H1, concluyendo que si hay diferencias en concentraciones de bacterias antes y después de aplicar el químico.

Pruebas de independencia.-

Las pruebas de independencia son aquellas en las cuales queremos determinar si dos caracteres o variables son independientes o no, esto es si existe alguna relación o independencia entre ambos.
La diferencia con los métodos de regresión y correlación que veremos mas adelante estriba en que aquí tratamos con caracteres cualitativos como sexo, existencia o no de una enfermedad, color especie, etc. o rangos de variables cuantitativas. y más que ver el punto exacto de correlación entre ambos vemos que cantidad de individuos poseen cierta combinación de caracteres y si esto es atribuible al azar o a una dependencia.

Los pasos para realizar esta prueba son los siguientes:

Enunciamos la hipótesis nula y su alterna:

H0 : Los caracteres son independientes.

H1 : Los caracteres son dependientes.

Construimos una tabla de doble clasificación, dividiéndola en s filas y r columnas, en donde r es el número de modalidades del carácter x y s es el número de modalidades del carácter y, poniendo en cada casilla fila por columna (ij), el número de observaciones que poseen el carácter x con la modalidad i y el carácter y con la modalidad j (nij):

Calculamos los valores esperados (npij) para cada casillero fila por columna (ij), mediante la fórmula:

$$np_{ij} = \frac{n_{i.}\, x\, n_{.j}}{N}$$

Calculamos el valor del estadístico $\chi^2$:

$$\chi^2 = \sum_{i=1}^{s} \sum_{j=1}^{r} \frac{(n_{ij} - np_{ij})^2}{np_{ij}}$$

el cual sigue una distribución Ji-cuadrado con n=(r-1)(s-1) grados de libertad.

La región de rechazo viene dada por:

W = $\{\chi^2 > \chi^2_{(\alpha)}\}$

Por ejemplo queremos saber si la circunferencia de la cabeza y la estatura son caracteres independientes en los niños recién nacidos, para lo cual disponemos de la siguiente tabla de contingencia:

**ESTATURA**

| Circunferen. | Pequeña | Mediana | Grande | Total |
|---|---|---|---|---|
| **Pequeña** | 40 | 36 | 2 | 78 |
| **Grande** | 0 | 14 | 7 | 21 |
| **Total** | 40 | 50 | 9 | 99 |

En donde x = circunferencia en cms. (<35=peq., >36=gran.) y y = estatura en cms (<49=peq, 50-52=med y >53=gran).

Calculamos los valores esperados multiplicando el valor de la suma de cada columna por el valor de la suma de la fila correspondiente y dividiéndola por el número total, como indica la fórmula.

| **Circunferen.** | Pequeña | Mediana | Grande | **Total** |
|---|---|---|---|---|
| **Pequeña** | 40x78/99= 31.5 | 50x78/99= 39.4 | 9x78/99= 7.1 | 78 |
| **Grande** | 40x21/99= 8.5 | 50x21/99= 10.6 | 9x21/99= 1.9 | 21 |
| Total | 40 | 50 | 9 | 99 |

Calculamos el valor de $\chi 2$ para cada casillero para facilitar y visualizar mejor los cálculos:

| **Circunferen.** | Pequeña | Mediana | Grande |
|---|---|---|---|
| **Pequeña** | $(40-31.5)^2/31.5$ = 2.29 | $(36-39.4)^2/39.4$ = 0.29 | $(2-7.1)^2/7.1$ = 3.66 |
| **Grande** | $(0-8.5)^2/8.5$ = 8.5 | $(14-10.6)^2/10.6$ = 1.09 | $(7-1.9)^2/1.9$ = 13.68 |

Y la suma será:

$$\chi^2 = 2.29 + 0.29 + 3.66 + 8.5 + 1.09 + 13.68 = 29.51$$

$$\chi^2 = 29.51$$

Buscando en la tabla encontramos que el valor de $\chi^2$ para $\alpha = 0.05$ y $\nu=(2-1)(3-1)=2$ grados de libertad es 5.991.
Y, dado que la región de rechazo:

$W = \{29.51 > 5.991\}$

Es verdadera, rechazamos H0 y aceptamos H1:
Los caracteres son dependientes.

<u>Prueba para comparar varias proporciones.-</u>

Otro uso para las tablas de contingencia es el determinar diferencias entre varias (k) proporciones.
Este es un caso especial de tablas de contingencia.
La principal diferencia es que tendremos dos renglones únicamente; uno con el número de éxitos y otro con el número de fracasos.
Aquí el número de grados de libertad será igual a n=k-1, siendo k el número de tratamientos.

Por ejemplo:

Los resultados de pruebas de stress de salinidad realizados en larvas de P. vannamei alimentados con distintas dietas constan a continuación:

| Postlarvas | Dieta 1 | Dieta 2 | Dieta 3 | Dieta 4 | Total |
|---|---|---|---|---|---|
| **Muertas** | 13 | 74 | 156 | 140 | 383 |
| **Vivas** | 437 | 376 | 294 | 160 | 267 |
| Total | 450 | 450 | 450 | 300 | 1,650 |

Queremos saber si los porcentajes de supervivencia al stress son los mismos o si difieren ($\alpha$=0.05).

H0: p1 = p2 = p3= p4
H1: al menos una pi no es igual

Los valores esperados serán:

| Postlarvas | Dieta 1 | Dieta 2 | Dieta 3 | Dieta 4 | Total |
|---|---|---|---|---|---|
| **Muertas** | 104.45 | 104.45 | 104.45 | 104.45 | 383 |
| **Vivas** | 343.55 | 343.55 | 343.55 | 343.55 | 267 |
| Total | 450 | 450 | 450 | 300 | 1,650 |

Calcule el valor de Ji- cuadrado respectivo y tome la decisión.

Análisis de Varianza

Al estudiar las pruebas de hipótesis, veíamos que éstas se usaban para comparar una o dos poblaciones.

En el caso de querer comparar más de dos poblaciones usando pruebas bimuestreales, tendríamos que determinar las diferencias entre cada par posible. Esto, además de ser tedioso, aumentaría la posibilidad total de un error hasta niveles prohibitivos.

Sin embargo, tenemos el análisis de varianza (ANDEVA o ANOVA), el cual no es otra cosa que una prueba de hipótesis para más de dos muestras, en donde tratamos de averiguar si las poblaciones que muestreamos son todas iguales en lo que respecta a sus medias, o si en su defecto, hay por lo menos una que sea distinta.

Matemáticamente el ANOVA de una vía para 2 niveles de un factor es equivalente a la prueba t bimuestreal.

Así, las hipótesis básicas que aplicaremos en un ANOVA serán:

Ho : $\mu1 = \mu2 = ... = \mu n$ (todas las medias son iguales)
H1 : $\exists$ $\mu i$ tal que $\mu \neq \mu i$ (al menos hay una media desigual)

Sin embargo, existen ANOVAs que no sólo comparan tratamientos, sino bloques; e incluso hay los que comparan varios factores y su efecto conjunto, por lo que estas hipótesis básicas pueden ser complementadas con otras, dependiendo del caso.

9.2.- ANOVA de una vía.-

Utilizamos un ANOVA de una vía cuando queremos comparar las medias en un experimento diseñado por azar simple.

Las hipótesis a probar son:

Ho : m1 = m2 = ... = mn (todas las medias son iguales)

H1 : $\exists$ $\mu i$ tal que $\mu \neq \mu i$ (al menos hay una media desigual)

Para facilidad en la realización de los cálculos, Al realizar ANOVA utilizamos las tablas de ANOVA, las cuales facilitan los cálculos

<u>ANOVA de una via.-</u>
Generalmente utilizamos un ANOVA de una vía cuando queremos comparar las medias en un experimento con diseño aleatorio simple. Este procedimiento descompone las fuentes de variaciones en la parte correspondiente al trataamiento y el del error.

Suponiendo que tenemos la siguiente tabla de datos para denptar nuestra simbología:

**TRATAMIENTOS**

| **1** | **2** | **....** | **j** | **....** | **k** | |
|---|---|---|---|---|---|---|
| $Y_{11}$ | $Y_{12}$ | ... | $Y_{1j}$ | ... | $Y_{1k}$ | |
| $Y_{12}$ | $Y_{22}$ | ... | $Y_{2j}$ | ... | $Y_{2k}$ | |
| ... | ... | ... | ... | ... | ... | |
| $Y_{i1}$ | $Y_{i2}$ | ... | $Y_{ij}$ | ... | $Y_{ik}$ | |
| ... | ... | ... | ... | ... | ... | |
| $Y_{n11}$ | $Y_{n22}$ | ... | $Y_{njj}$ | ... | $Y_{nkk}$ | **Totales** |
| $\Sigma Y_{i1}$ | $\Sigma Y_{i2}$ | ... | $\Sigma Y_{i\,j}$ | ... | $\Sigma Y_{i\,k}$ | $\Sigma\Sigma Y_{ij}$ |

| | | | | | | |
|---|---|---|---|---|---|---|
| $n_1$ | $n_2$ | ... | $n_j$ | ... | $n_k$ | N |
| $\Sigma Y^2_{i1}$ | $\Sigma Y^2_{i\,2}$ | ... | $\Sigma Y^2_{ij}$ | ... | $\Sigma Y^2_{i\,k}$ | $\Sigma\Sigma Y^2_{ij}$ |

La tabla de Anova para los cálculos será la siguiente:

| **Fuente de Variación** | **Gdos Libertad** | **Suma de Cuadrados (SC)** | **Sum med Cuadrados (SMC)** | ***F*** |
|---|---|---|---|---|
| Tratamiento | k-1 | $\frac{\sum_{j=1}^{k}(\sum_{i=1}^{n} Y_{ij})^2}{nj} - C$ | SCT/(k-1) | SMCT/SMCE |
| Error | N-k | SC Tot - SCT | SCE/(N-k) | |
| Total | N-1 | $\sum_{j=1}^{k}\sum_{i=1}^{n} Y_{ij}^2 - C$ | | |

En donde C es igual a :

$$C = \frac{(\sum_{j=1}^{k}\sum_{i=1}^{n} Y_{ij})^2}{N}$$

Y, en donde la relación **SMCT/SMCE** es nuestro estadístico de prueba *F*, el cual sigue una distribución “F” de Fisher con $\nu_1$ = k-1 y $\nu_2$ = N-k grados de libertad.

La región crítica o de rechazo viene dada por:

$$\boldsymbol{F} \geq \mathbf{F}_{(\alpha)}$$

Veamos un ejemplo de este ANOVA:

Los siguientes son los datos de crecimiento semanal de camarón en 19 piscinas divididas en 4 tratamientos que representan 4 dietas distintas:

**TRATAMIENTOS**

| **1** | **2** | **3** | **4** | |
|---|---|---|---|---|
| 0.74 | 0.68 | 0.75 | 0.72 | |
| 0.76 | 0.71 | 0.77 | 0.74 | |
| 0.75 | 0.71 | 0.77 | 0.73 | |
| 0.74 | 0.72 | 0.76 | 0.73 | |
| 0.75 | | 0.75 | | |
| 0.76 | | | | **Totales** |

| | | | | | |
|---|---|---|---|---|---|
| $\Sigma Y_{ij}$ | 4.50 | 2.82 | 3.80 | 2.92 | 14.04 |
| $n_j$ | 6 | 4 | 5 | 4 | 19 |
| | 3.38 | 1.99 | 2.89 | 2.13 | 10.384 |
| $\Sigma Y^2_{ij}$ | 0.75 | 0.71 | 0.76 | 0.73 | 0.74 |

Deseamos saber si hay diferencias entre tratamientos con respecto a sus medias (a=0.05).

Calculamos la suma de cuadrados para los tratamientos:

$$SCT = \frac{4.5^2}{6} + \frac{2.82^2}{4} + \frac{3.8^2}{5} + \frac{2.92^2}{4} - \frac{14.04^2}{19} =$$

SCT = 3.375 + 1.9881 + 2.888 + 2.1316 - 10.374821 = 0.0078789

Calculamos la suma de cuadrados totales :

$$SCTot = 10.3846 - \frac{14.04^2}{19} =$$

SCTotal = 10.3846 - 10.374821 = 0.0097789

Construimos nuestra tabla de ANOVA:

| **Fuente de Variación** | **Gdos Libertad** | **Suma de Cuadrados (SC)** | **Sum med Cuadrados (SMC)** | ***F*** |
|---|---|---|---|---|
| Dietas | 3 | 0.0078789 | 0.0026263 | 20.74 |
| Error | 15 | 0.0019 | 0.0001266 | |
| Total | 18 | 0.0097789 | | |

Buscamos en la tabla F el valor de F0.05 para $\nu 1=3$ y $\nu 2=15$ grados de libertad = 3.29.

Y como:

W = { 20.74 > 3.29 }

Rechazamos H0, aceptamos H1 y concluimos que hay diferencias significativas entre las medias ($\alpha$=0.05).

ANOVA de dos vías.-

Utilizamos un ANOVA de dos vías cuando queremos comparar las medias de un experimento con dos criterios de clasificación como, por ejemplo, con un diseño por bloques aleatorios; es decir, cuando tenemos otra fuente de variación conocida, además de nuestro tratamiento.

De esta forma tomamos en cuenta el error que podría haber al considerar iguales a dos poblaciones con el mismo tratamiento, pero realizado en distintas condiciones.

Se llama de dos vias, porque a diferencia del anterior, en este separamos 2 fuentes de variación además de la del error.

También lo podemos usar para probar muestras dependientes como, por ejemplo, las pruebas de antes y después, considerando cada muestra pareada como un bloque y antes y después como los tratamientos.

Las hipótesis básicas a probar son:

Ho : $\mu 1 = \mu 2 = ... = \mu n$ ( Todas las $\mu$ de los trat. son iguales)

H1 : $\exists$ $\mu i$ tal que $\mu \neq \mu i$ (existe al menos una $\mu$ de los trat. desigual)

Es decir, probar los efectos de los tratamientos.

Pudiéndose también probar:

Ho' : $\mu a = \mu b = ... = \mu k$ ( todas las $\mu$ de los bloques son iguales)

H1' : $\exists$ $\mu j$ tal que $\mu \neq \mu j$ (existe ak menos una $\mu$ de los bloques desigual)

O sea, probar los efectos de los bloques.

Las tablas usadas en este caso son las siguientes:

**TRATAMIENTO**

| **BLOQUE** | **A** | **B** | **....** | **i** | **....** | **n** | **Ti.** |
|---|---|---|---|---|---|---|---|
| **1** | $Y_{A1}$ | $Y_{B1}$ | ... | $Y_{i1}$ | ... | $Y_{n1}$ | $\Sigma^n Y_{i1}$ |
| **2** | $Y_{A2}$ | $Y_{B2}$ | ... | $Y_{i2}$ | ... | $Y_{n2}$ | $\Sigma^n Y_{i2}$ |
| **...** | ... | ... | ... | ... | ... | ... | ... |
| **j** | $Y_{Aj}$ | $Y_{Bj}$ | ... | $Y_{ij}$ | ... | $Y_{nj}$ | $\Sigma^n Y_{i\,j}$ |

| ... | ... | ... | ... | ... | ... | ... | ... |
|---|---|---|---|---|---|---|---|
| **k** | $Y_{Ak}$ | $Y_{Bk}$ | ... | $Y_{ik}$ | ... | $Y_{nk}$ | $\Sigma^n Y_{i\,k}$ |
| **T.j** | $\Sigma^k Y_{Aj}$ | $\Sigma^k Y_{Bj}$ | ... | $\Sigma^k Y_{ij}$ | ... | $\Sigma^k Y_{nj}$ | $\Sigma^k \Sigma^n Y_{ij}$ |
| **$T^2$.j** | $\Sigma^k Y^2_{A1}$ | $\Sigma^k Y^2_{B1}$ | ... | $\Sigma^k Y^2_{i1}$ | ... | $\Sigma^k Y^2_{n1}$ | $\Sigma^k \Sigma^n Y^2_{A1}$ |

En donde **n** es el número de tratamientos, **k** el n el no de bloques, y, **N** (nxk) es el número total de observaciones.
La tabla de ANOVA usada es la siguiente:

| **Fuente de Variación** | **Gdos Libertad** | **Suma de Cuadrados (SC)** | **Sum med Cuadrados (SMC)** | ***F*** |
|---|---|---|---|---|
| Tratamiento | n-1 | $\frac{\sum_{i=1}^{n}(\sum_{j=1}^{k} Y_{ij})^2}{k} - C$ | SCT/(n-1) | SMCT/SMCE |
| Bloques | k-1 | $\frac{\sum_{i=1}^{k}(\sum_{j=1}^{n} Y_{ij})^2}{n} - C$ | SCB/(k-1) | SMCB/SMCE |
| Error | (n-1)(k-1) | SC Tot - SCT | SCE/(n-1)(k-1) | |
| Total | N-1 | $\sum_{j=1}^{k}\sum_{i=1}^{n} Y_{ij}^2 - C$ | | |

En donde C es igual a :

$$C = \frac{(\sum_{j=1}^{k}\sum_{i=1}^{n} Y_{ij})^2}{N}$$

Y, en donde la relación **SMCT/SMCE** es nuestro estadístico de prueba $F_1$ para tratamientos, el cual sigue una distribución "F" de Fisher con $\nu_1$ = n-1 y $\nu_2$ = (n-1)(k-1 grados de libertad.

La región crítica o de rechazo viene dada por:

$$\boldsymbol{F} \geq \mathbf{F}_{(\alpha)}$$

La r elación **SMCB/SMCE** es nuestro estadístico de prueba $F_1$ para bloques, el cual sigue una distribución "F" de Fisher con $\nu_1$ = k-1 y $\nu_2$ = (n-1)(k-1) grados de libertad.

La región crítica o de rechazo viene dada por:

$$\boldsymbol{F} \geq \mathbf{F}_{(\alpha)}$$

Veamos un ejemplo:
(Villalón, 92)
Se quería probar la respuesta de P. vannamei a dos tipos de harina de trigo en el balanceado. Para esto, se tomaron 6 precriaderos que estaban disponibles en 3 camaroneras situadas en distintas zonas de la provincia del Guayas. Esto se lo hizo por dos razones. Una porque no habían disponibles 6 precriaderos en una sola camaronera, y otra porque de esta forma los resultados no serían exclusivos para la camaronera en la cual se los realizó. El sistema de cultivo y procedencia de las larvas fue idéntica para todas las piscinas. siendo la única diferencia por nosotros aplicada el tipo de harina de trigo. La otra diferencia detectada pero que no pudimos (o no quisimos) controlar fue la ubicación.

Para evaluar estos datos se optó por un diseño en bloques aleatorios, en el cual cada camaronera representaba un bloque donde se pensaba que iba a aparecer una diferencia de crecimiento entre ellas.

Los siguientes son los datos de incremento semanal promedio de
peso:

| **Bloque** | **FI** | **FII** | **Ti.** |
|---|---|---|---|
| Bonanza | 0.79 | 0.62 | 1.41 |
| Fafra | 1.08 | 0.85 | 1.93 |
| Toyo | 0.67 | 0.72 | 1.39 |
| **T.j** | 2.54 | 2.19 | 4.73 |
| **$T^2$.j** | 2.2394 | 1.6253 | 3.8647 |

| | | | |
|---|---|---|---|
| C= | 3.728816 | n= | 2 |
| SS total= | 0.14 | k= | 3 |
| SS treat= | 0.02 | N= | 6 |
| SS block= | 0.09 | | |
| SS error= | 0.02 | | |

ANOVA TABLE

| SOURCE | df | SS | MS | F |
|---|---|---|---|---|
| TREATMENT | 1 | 0.02 | 0.02 | 1.878834 |
| BLOCK | 2 | 0.09 | 0.05 | 4.312883 |
| ERROR | 2 | 0.02 | 0.01 | |
| TOTAL | 5 | 0.14 | 0.03 | |

BETWEEN TREATMENTS

H0= u1=u2
H1= reject H0

| | |
|---|---|
| alfa | 0.05 |
| df. num | 1 |
| df. den | 2 |
| F calc | 1.88 |
| F tabla | 18.50 |

So : u1=u2

BETWEEN BLOCKS

H0= u1=u2=u3
H1= reject H0

| | |
|---|---|
| alfa | 0.05 |
| df. num | 2 |
| df. den | 2 |
| F calc | 4.31 |
| F tabla | 19.00 |

So : u1=u2=u3

Para usarlo en muestras dependientes tendríamos el ejemplo que usamos para la prueba "t":

Se deseaba determinar la eficiencia de cierto químico como bactericida en tanques de larvas, para lo cual se realizan los siguientes contajes de bacterias en el agua de 12 tanques antes y después de su aplicación:

| ANTES | DESPUES |
|---|---|
| 15.200 | 16.300 |
| 10.100 | 8.300 |
| 6.400 | 28.100 |
| 4.300 | 6.800 |
| 9.300 | 11.400 |
| 12.800 | 34.000 |
| 8.400 | 9.300 |
| 6.500 | 4.200 |
| 3.100 | 12.300 |
| 2.400 | 4.500 |
| 9.400 | 4.200 |
| 6.300 | 11.100 |

Determine si hay diferencias significativas (a = 0.01) debidas al tratamiento usando un ANOVA de dos vias.

Restricciones

Los análisis de varianza se basan en ciertas suposiciones que debemos de tomar en cuenta al realizarlos. Estas son:

- El proceso está bajo control y puede ser repetido.
- Todas las muestras deben ser aleatorias e independientes.
- Las poblaciones de las cuales muestreamos deben de seguir una distribución Normal.
- Todas las poblaciones que estamos probando deben de tener varianzas iguales.

Algunos libros discuten estas suposiciones y qué se puede hacer si éstas no se cumplen.

Algunos estudios han demostrado que la falta de normalidad no afecta seriamente el análisis cuando el tamaño de la muestra es igual para todos los tratamientos.

En el caso de que la varianza y la media sean dependientes entre sí, existen tablas que indican cómo romper esta dependencia; sin embargo este caso no será estudiado por nosotros.

Un caso muy común para nosotros puede ser el querer comparar diferentes pH. Para este caso por ejemplo, lo que debemos hacer es transformar los valores de pH en concentraciones Molares de Hidrógeno mediante el uso de la fórmula $10^{-pH}$ y procesar estos datos para ver si hay diferencias entre [H].

Para probar la igualdad de varianzas podemos usar varias pruebas. Una forma fácil de hacerlo es observar el rango de cada tratamiento. Se saca un rango promedio , el cual se lo multiplica por un factor D4 extraído de la tabla anexa para distintos tamaños de muestra. Si el productor D4 es mayor que todos los rangos, se puede asumir con alguna seguridad que las varianzas son homogéneas.

Para el ejemplo de ANOVA ya revisado:

| 1 | 2 | 3 | 4 |
|---|---|---|---|
| 0.74 | 0.68 | 0.75 | 0.72 |
| 0.76 | 0.71 | 0.77 | 0.74 |

| | | | | | |
|---|---|---|---|---|---|
| | 0.75 | 0.71 | 0.77 | 0.73 | |
| | 0.74 | 0.72 | 0.76 | 0.73 | |
| | 0.75 | | 0.75 | | |
| | 0.76 | | | | **Totales** |
| $\Sigma Y_{ij}$ | 4.50 | 2.82 | 3.80 | 2.92 | 14.04 |
| $n_j$ | 6 | 4 | 5 | 4 | 19 |
| | 3.38 | 1.99 | 2.89 | 2.13 | 10.384 |
| $\Sigma Y^2_{ij}$ | 0.75 | 0.71 | 0.76 | 0.73 | 0.74 |

Los rangos serán los siguientes:

A = 0.76-0.74=0.02
B = 0.72-0.68=0.04
C = 0.77-0.75=0.02
D = 0.74-0.72=0.02

Y el rango medio =(0.02+0.04+0.02+0.02)/4=0.025

El valor de la tabla para 4 tratamientos es :

D4= 2.282

Entonces
rD4 = 0.025 x 2.282 = 0.057

Y, como:

0.057 > 0.02 es V
0.057 > 0.04 es V
0.057 > 0.02 es V
0.057 > 0.02 es V

Podemos decir que se tratan de varianzas homogéneas.

Otras pruebas existentes son la prueba C de Cochran y la prueba Fmax de Hartley, sin embargo estas pruebas solo sirven para igual número de repeticiones. La mayoría de programas estadísticos cuentan con herramientas para evaluar la homoceisdad de las muestras.

Para la prueba De Cochran dividimos la varianza máxima observada para la suma de todas las varianzas observadas;

C = s2 max / S s2

Y si el valor obtenido es menor que el valor del nomógrafo de la tabla correspondiente, entonces podemos decir que las varianzas son homogéneas.

Para la prueba Fmax de Hartley, simplemente dividimos la mayor varianza observada por la menor varianza observada:

Fmax= s2max / s2min

y comparamos con el valor de la tabla adjunta.

<u>Comparaciones múltiples.-</u>

Después de realizar nuestro ANOVA, sabremos si todos nuestros tratamientos son iguales o si existe al menos uno que sea distinto; pero, usualmente queremos saber más: Cuál tratamiento es mejor y cuál es peor? Cuál tratamiento difiere de los otros y cuáles son iguales entre sí?

Existen muchas formas de responder estas preguntas, entre ellas tenemos las pruebas de los contrastes ortogonales, la prueba de Scheffé la prueba de Tukey, la prueba de rango múltiple de Duncan y la prueba de Student Newman Keuls o S-N-K. Aquí sólo estudiaremos esta última.

<u>Prueba Student- Newman-Keuls (SNK).-</u>

Es recomendada para ser usada cuando la decisión de cuál comparación debe hacerse se la realiza después de que los datos han sido examinados.
Las comparaciones se hacen mediante amplitudes múltiples basándose en los resultados.

Veamos un ejemplo con el ejercicio anterior.
Los pasos a seuir son:

1.- Ordene los k medias de menor a mayor.

| **N** | **6** | **4** | **5** | **4** |
|---|---|---|---|---|
| **Tratam.** | B | D | A | C |
| **x** | 0.71 | 0.73 | 0.75 | 0.76 |

2.- Determine el MSCE y sus grados de libertad de la tabla de ANOVA.

Revisamos la tabla de ANOVA, en donde la SMCE era 0.0001266 con $\nu$=15 grados de libertad.

3.- Obtenga el error estándar de la media para cada tratamiento mediante la fórmula:

$$Sx_j = \sqrt{\frac{MSCE}{n_j}}$$

Aquí encontramos un problema, el número de réplicas no es el mismo, pero con cierto grado de aproximación se permite remplazar nj por la media de los n de los dos tratamientos que compararemos, por lo que en este caso no calcularemos sxj todavía, ya que no va a ser el mismo para todas las comparaciones que hagamos.

4.- Busque en la tabla de rangos studentizada los valores de los rangos significativos (rp) a un nivel a, con $\nu 2$ grados de libertad de la tabla de ANOVA y
p = 2,3,4, y apunte estos k-1=4-1 = 3 rangos.

Buscando en la tabla tenemos:

| | | | |
|---|---|---|---|
| p | 2 | 3 | 4 |
| rp | 3.01 | 3.67 | 4.08 |

5.- Multiplique estos rangos por Sxj para lograr un grupo de k-1 rangos menos significativos.

Como todos los Sxj no so n iguales seguimos al siguiente punto.

6.- Pruebe todas las k(k-1)/2 posibles diferencias entre pares de medias con sus respectivos valores de rangos menos significativos.

Los valores de diferencias ³ que el rango de mínima significancia indican diferencias significativas.

Aquí se comparan 4(4-1)/2= 6 diferencias.

Comparamos B con C:

3 espacios de distancia (p=4)

rp = 4.08

diferencia = 0.76 - 0.71 =0.05

$$Sx_j = \sqrt{\frac{0.0001266}{(6+4)/2}}$$

Sy=0.0050318

Rp=0.0050318 x 4.08 = 0.02

0.05 > 0.02 Verdadero entonces diferentes

Comparamos B con A:

2 espacios de distancia (p=3)

rp = 3.67

diferencia = 0.75 - 0.71 =0.04

$$Sx_j = \sqrt{\frac{0.0001266}{(6+5)/2}}$$

Sy=0.0047977

Rp=0.0047977 x 3.67 = 0.017

0.04 > 0.017 verdadero entonces diferentes

Comparar B con D:

1 espacios de distancia (p=2)

rp = 3.01

diferencia = 0.73 - 0.71 =0.02

$$Sx_j = \sqrt{\frac{0.0001266}{(6+4)/2}}$$

Sy=0.0050318

Rp=0.0050318 x 3.01 = 0.015

0.02 > 0.015 verdadero entonces diferentes

Comparar D con C:

2 espacios de distancia (p=3)

rp = 3.67

diferencia = 0.76 - 0.73 =0.03

$$Sx_j = \sqrt{\frac{0.0001266}{4}}$$

Sy=0.0056258

Rp=0.0056258 x 3.67 = 0.02

0.03 > 0.02 verdadero entonces diferentes

Comparar D con A:

1 espacios de distancia (p=2)

rp = 3.01

diferencia = 0.73 - 0.75 =0.02

$$Sx_j = \sqrt{\frac{0.0001266}{(4+5)/2}}$$

Sy=0.005304

Rp=0.005304 x 3.01 = 0.015

0.02 > 0.015 verdadero entonces diferentes

Comparar A con C:

1 espacios de distancia (p=2)

rp = 3.01

diferencia = 0.76 - 0.75 =0.01

$$Sx_j = \sqrt{\frac{0.0001266}{(5+4)/2}}$$

Sy=0.005304

Rp=0.005304 x 3.01 = 0.0159

0.01 > 0.0159 Falso entonces iguales.

Para expresar los resultados veremos dos formas.

La manera sugerida por Duncan es colocar los valores de los promedios a distancias equivalentes a sus valores y unir los promedios homogéneos mediante una línea:

```
  B     D     A    C
0.71  0.73  0.75 0.76
            -------------
```

Otra manera es colocar subíndices iguales a las medias homogéneas:

A 0.75 (c)
B 0.71 (a)
C 0.76 (c)
D 0.73 (b)

<u>ANOVA multifactorial.-</u>

Utilizamos un ANOVA multifactorial cuando queremos comparar los efectos de mas de 1 tratamiento, así como los efectos de las interacciones de esos tratamientos.
Es decir, cuando tenemos un diseño factorial.

En este capítulo veremos como se realiza un ANOVA de 2 factores.

Definiremos $\alpha$ como el efecto del factor # 1, $\beta$ como el efecto del factor # 2, $\rho$ como el efecto de las repeticiones y $(\alpha\beta)$ como el efecto de las interacciones de ambos factores.

Las hipótesis a probar son:

H0 : $\alpha 1=\alpha 2=...=\alpha=0$ (todos los efectos del 1er factor son 0)

H1 : $\exists$ $\alpha i$ tal que $ai\neq 0$ (al menos 1 efecto del 1er factor no es 0)

Pudiéndose probar también:

H0' : $\beta 1=\beta 2=...=\beta=0$ (todos los efectos del 2do factor son 0)

H1' : $\exists$ $\beta j$ tal que $\beta j\neq 0$ (al menos 1 efecto del 2do factor no es 0)

y:

H0'' :$\rho r1=\rho 2=...=\rho=0$ (todos los efectos de las réplicas son 0)

H1'' : $\exists$ $\rho k$ tal que $\rho k\neq 0$ (al menos 1 efecto de las réplicas no es 0) así como:

H0''' : Todos los $(\alpha\beta)ij$ son iguales.

H1''' : Al menos un $(\alpha\beta)ij$ no es igual.

No veremos la mecánica de esta prueba debido a lo corto del tiempo.

Uso de la regresión lineal para comparaciones de dos variables cuantitativas.-

Cuando estamos comparando datos de dos variables cuantitativas y nuestro ANOVA nos da diferencias muy significativas, podemos recurrir a la regresión lineal para determinar cómo responde nuestra variable dependiente y a las variaciones dentro de un rango de x.

El primer paso a seguir es calcular la ecuación lineal con sus constantes **a** y **b**. Con base en esto elaboramos una tabla con los valores de **x**, **y~**, **y'** (calculada a partir de la ecuación para cada valor de x), **Δy** (y~ - y') y **(Δy)$^2$**.

Con estos datos elaboramos una nueva tabla de ANOVA:

| **Fuente de Variación** | **Gdos Libertad** | **Suma de Cuadrados (SC)** | **Sum med Cuadrados (SMC)** | ***F*** |
|---|---|---|---|---|
| Tratamiento | k-1 | $\sum_{j=1}^{n}\left(\frac{\left(\sum_{j=1}^{k} Y_{ij}\right)^2}{n_j}\right) - C$ | SCT/(k-1) | SMCT/SMCE |
| Lineal | 1 | SCT-SCDL | SCL/1 | SMCL/SMCE |
| Desviación de Lineal | k-2 | $n_j \ (\Delta y)^2$ | SCDL/(k-2) | SMCDL/SMCE |
| Error | N-k | SC Tot - SCT | SCE/N-k | |
| Total | N-1 | $\sum_{j=1}^{k}\sum_{i=1}^{n} Y_{ij}^2 - C$ | | |

En donde las hipótesis a probar son:

H0: No hay diferencias atribuibles a regresión lineal
H1: Si hay diferencias atribuibles a regresión lineal

Con $F$ = SMCL/SMCE con $\nu 1 = 1$ y $\nu 2$ = N-k grados de libertad como estadístico de prueba.

H0' : No hay desviaciones de la regresión lineal calculada
H1' : Si hay desviaciones de la regresión lineal calculada

Con $F$ = SMCDL/SMCE con $\nu$1= k-2 y $\nu$2 = N-k grados de libertad como estadístico de prueba.

Veamos un ejemplo:

Los siguientes datos corresponden a los tiempos de reacción de un pez a diferentes concentraciones de cierta droga:

| | Tratamientos Dosis (x) | | | | | |
|---|---|---|---|---|---|---|
| | 0.5 | 1.0 | 1.5 | 2.0 | 2.5 | |
| | 26.0 | 28.1 | 29.2 | 34.0 | 34.9 | |
| | 26.1 | 29.1 | 31.4 | 32.5 | 35.8 | |
| | 26.2 | 28.6 | 31.6 | 34.6 | 35.6 | Total |
| y | 26.1 | 28.6 | 30.73 | 33.7 | 35.43 | 30.91 |
| n | 3 | 3 | 3 | 3 | 3 | 15 |
| Sy2 | 2043.7 | 2454.4 | 2837.2 | 3409.4 | 3767.0 | 14511.61 |
| Sy | 78.3 | 85.8 | 92.2 | 101.1 | 106.3 | 463.7 |
| (Sy)2/n | 2043.63 | 2453.88 | 2833.61 | 3407.07 | 3766.56 | 14504.75 |

Primero realizamos el ANOVA:

C=463.72/15=14,334.51
SCTOT=SSY2ij-C=14511.61-14334.51=177.1
SCT=S(Sy)2/n-C=14504.75-14334.51=170.24
SCE=177.1-170.24=6.86

| FUENTE | GDL | S.C | S.M.C. | F |
|---|---|---|---|---|
| TRATAMIENTO | 4 | 170.24 | 42.56 | 62.04 |
| ERROR | 10 | 6.86 | 0.686 | |
| TOTAL | 14 | 177.1 | | |

Y como 62.04>>5.96, rechazamos H0 y concluimos que hay diferencias altamente significativas.

El siguiente paso a seguir es calcular la ecuación lineal con sus constantes a y b.

731.2 - (463.7)(22.5)/15

$$b = \frac{\qquad}{41.25 - (22.5)^2 / 15} = 4.75$$

a = 463.7 / 15 - 4.75 (22.5 / 15) = 23.79

Siendo la ecuación correspondiente:

y = 23.79 + 4.75 x

Con base en esto elaboramos una tabla con los valores de x, y, y'(calculada a partir de la ecuación para cada valor de x),Dy (y - y') y (Dy)2.

| **x** | y | y' | $\Delta$y | **($\Delta$y)$^2$** |
|---|---|---|---|---|
| **0.5** | 26.10 | 26.165 | -0.065 | 0.004225 |
| **1.0** | 28.60 | 28.540 | -0.060 | 0.0036 |
| **1.5** | 30.733 | 30.915 | -0.182 | 0.03314 |
| **2.0** | 33.70 | 33.290 | -0.410 | 0.1681 |
| **2.5** | 35.433 | 33.665 | -0.232 | 0.053824 |
| **Total** | | | | 0.262873 |

Con base en los datos de nuestra tabla de x y y, elaboramos nuestra nueva tabla de ANOVA:

ANOVA TABLE

| SOURCE | df | SS | MS | F |
|---|---|---|---|---|
| TREATMENT | 4 | 170.24 | 42.56 | 61.24 |
| Lineal | 1 | 169.45 | 169.45 | 243.81 |
| Desv Lineal | 3 | 0.7886 | 0.26287 | 0.38 |
| ERROR | 10 | 6.95 | 0.695 | |
| TOTAL | 14 | 177.19 | | |

SCDI = 3 x 0.262873 = 0.7886
SCL = 170.24 - 0.7886 = 169.45

En donde las hipótesis a probar son:
Ho : No hay diferencias atribuibles a regresión lineal.
H1 : Sí hay diferencias atribuibles a regresión lineal.

Para el estadístico F = SMCL / SMCE con n1 = 1 y n2 = N-k grados de libertad.

y:

Ho : No hay desviaciones de la regresión lineal calculada.
H1 : Si hay desviaciones de la regresión lineal calculada.

Para el estadístico F = SMCD / SMCE con n1 = k-2 y n2 = N-k grados de libertad.

**Diseño Experimental**

Llamamos investigación a la "búsqueda sistemática de la verdad no descubierta" (Leedy, 1974), poniendo especial atención a la palabra "sistemática".

Un experimento puede ser definido como "un estudio donde cierta(s) variable(s) independiente(s) es(son) manipulada(s), y su(s) efecto(s) en una o más variables independiente(s) determinado(s), siendo los niveles de esta(s) variable(s) independientes asignados al azar" (Hicks, 1982).

El experimento debe incluir una enunciación especifica del problema que se quiere solucionar, o sea que se debe describir exactamente lo que se quiere averiguar.
Esto parecerá bastante obvio, pero en la práctica toma un tiempo mas o menos llegar a un acuerdo general sobre cual el problema específico es.
Un buen enunciado que tome en cuenta todos los puntos da una gran ventaja en la realización de un experimento.

No es bueno enunciar el problema en forma general, sino más bien en una forma en que todos los puntos queden claros y que nos muestren la forma en que la investigación debe ser llevada.

El enunciado del problema debe incluir la referencia a uno o más criterios (variables dependientes) usados para determinar el problema. Se debe determinar si estos criterios pueden ser medidos, con cuánta precisión y con qué medios.

El enunciado también debe definir los factores (variables independientes) que afectarán a estos criterios y si éstos van a ser mantenidos fijos o variados a ciertos niveles o al azar, el número de factores , el tipo y su disposición, y también si estos factores van a ser combinados y en qué forma, todo lo cual determinará el tipo de cálculo estadístico a realizar.

Un ejemplo clásico en el diseño de experimentos es el realizado para determinar la eficacia de la vacuna Salk para polio.
Aquí la pregunta inicial fue "Que puede hacer la vacuna Salk al polio?". Como todos podemos ver este es el tipo de pregunta que no queremos tener porque es difícil o imposible de responder. La pregunta final a la cual se le dio respuesta fue:
"Hay una diferencia en el porcentaje de alumnos de 1o, 2o y 3o grado en los EE.UU. que contraen polio durante el primer año después de haber sido inoculados con la vacuna Salk y aquellos que no fueron vacunados?
Como vemos, aquí casi todo el trabajo está hecho.

Los factores (variables independientes) básicamente pueden ser manejados así:

- Controlados rígidamente y mantenidos fijos a lo largo del experimento, con lo cual los resultados obtenidos son válidos solamente para estas condiciones fijas. Por ejemplo en la vacuna Salk, decidimos estudiar solo los 3 primeros grados, ya que era en niños de esa edad donde mas prevalente era esa enfermedad.

Otro ejemplo si queremos probar cuatro alimentos para larvas y los factores de densidad, manejo de agua ,mallas antibióticos, origen de larva, etc. los mantenemos fijos entre los cuatro tratamientos según nos interese a nosotros. De esta forma las diferencias obtenidas en los resultados se deberían al tratamiento o factor que nos interesa (alimentación), ya que sería la única diferencia entre los tanques.

- Controlados a niveles fijos de interés.

Este es el caso de la(s) variable(s) que nos interesa(n), y que pueden ser cuantitativas o cualitativas.
Por ejemplo cuantitativas serían diferentes dosis de droga aplicadas a un animal o diferentes porcentajes de proteína en un alimento.
Cualitativos serían: La aplicación o no de la vacuna, Diferentes marcas de alimento para larvas, etc.

Resulta conveniente a veces incluir los extremos de esta(s) variable(s) para maximizar el efecto si lo hubiera. Otro punto es incluir los valores de esta variable que mas nos interesen.

- Aleatorios para promediar el efecto de variables que no pueden ser controladas.
El orden de experimentación es aleatorizado para promediar los efectos de ciertas variables que desconocemos o que no podemos controlar.
Un ejemplo de esto sería en el experimento de los distintos alimentos para larvas el aleatorizar la asignación del tratamiento a cada tanque, la aleatorización del orden de siembra, etc.

Este promedio sin embargo no remueve completamente el efecto de estas variables, ya que estas siguen incrementando la varianza de los datos observados. Sin embargo, el correcto diseño y planificación pueden reducir o minimizar el error anticipándose a estos factores.
Este es el caso si queremos probar dos tipos de balanceado en tres camaroneras. Nosotros podremos anticipar el efecto del factor camaronera, minimizándolo mediante un buen diseño.

Un principio básico es maximizar el efecto de la(s) variable(s) de interés, minimizar la varianza del error y controlar ciertas variables a niveles específicos.

El diseño del experimento involucra la cantidad de observaciones que se van a realizar, el orden en el cual se va a efectuar (que debería ser aleatorio para promediar las diferencias de ciertas variables que no podemos controlar, así como para poder asumir que los errores en medición son independientes), y también el orden de aleatoriedad a usar.

Al haber decidido esto se debe de mantener lo más apegado posible al plan trazado.

Debemos además estar al tanto de las restricciones del modelo que estamos usando y expresar el problema en forma de una hipótesis que podamos probar, y su hipótesis alterna.(todo esto ya lo vimos)

Debe de tomarse en cuenta en el diseño de forma muy especial las restricciones de tipo logístico y financieras, ya que habrá que hacer un compromiso entre estas y lo óptimo en términos matemáticos.

El análisis de datos es el último paso e incluye la realización mecánica de pasos ya decididos como son:

- Recolección y procesamiento de datos.

- Cálculo de ciertos estadísticos de prueba usados para hacer una decisión.

- La toma de la decisión en sí.

- La exposición de los resultados.

Los resultados deben ser expuestos de forma clara, precisa e inteligible, y los cálculos estadísticos pueden ser incluidos en los apéndices, sólo de ser necesario.

Repeticiones.-

Las repeticiones o replicas son importantes por:

- Permitir una estimación del error experimental.
- Mejorar la precisión de un experimento mediante la reducción de la varianza de la media.
- Aumentar el alcance de las inferencias de un experimento a través de la selección y del uso apropiado de unidades experimentales (a más tipos de condiciones del mismo tratamiento).

Modelos de diseño.-

Existen innumerables modelos de diseño de experimentos, no siendo el propósito de este manual el abarcarlos todos; sin embargo se explicarán algunos modelos de diseños básicos, lo que facilitará la comprensión de otros diseños que se encuentren en otros libros más avanzados.

Diseño de un factor, completamente aleatorio.-

Siempre que sólo se varíe un factor, ya sea cualitativo o cuantitativo, fijo o al azar, el experimento se considera de un factor.

Dentro de este factor existirán varios niveles o tratamientos, cuyo número lo designaremos como k. El objetivo del experimento es determinar si existen diferencias entre los efectos de los tratamientos.

Si el orden de experimentación aplicado a los distintos tratamientos es completamente aleatorio, de forma que se consideren aproximadamente homogéneas las condiciones en que se está trabajando, llamaremos a este diseño completamente aleatorio.

Cada tratamiento tendrá ni observaciones, sin importar que el tamaño de las muestras no sea igual de un tratamiento a otro.

El modelo matemático vendrá dado por:

$$Y_{ij} = \mu + \tau_j + \varepsilon_{ij}$$

Lo que significa que cada iésima observación del jésimo tratamiento va a ser igual a la media poblacional (m) más un efecto del jésimo tratamiento (tj) mas un error aleatorio para el mismo (e ij).

Este diseño se lo resuelve generalmente mediante un ANOVA de una vía con la hipótesis nula:

H0 : tj = 0 ; para todos los j.

Diseño de un factor, bloques aleatorios.-

En el caso de que no todas las observaciones que podamos realizar para los n tratamientos puedan considerarse homogéneas, y es más, podamos considerar que hay fuentes conocidas de variabilidad, nos podemos librar de esta variabilidad dividiendo las observaciones de cada clasificación en bloques.

En el caso de que en cada bloque haya una observación de cada tratamiento, y siempre y cuando los tratamientos sean asignados al azar dentro de cada bloque, denominamos a este diseño en bloques aleatorios.

El modelo matemático usado vendrá dado por:

$$Y_{ij} = \mu + \beta_j + \tau_i + \varepsilon_{ij}$$

en donde bj es el efecto del jésimo bloque y ti es el efecto del iésimo tratamiento.

Este diseño se lo resuelve mediante un ANOVA de dos vías, en donde la hipótesis fundamental a probar es:

H0 : ti = 0 , para todos los tratamientos,

pudiéndose también probar:

H0 : bj = 0 , para todos los bloques,

la cual, al rechazarse, nos indicará que el criterio de clasificación en bloques ha sido correcto.

<u>Cuadrado Latino.-</u>

En el caso de que queramos eliminar dos fuentes de variación conocidas, en un experimento de un solo factor, utilizaremos el diseño de cuadrado latino, el cual es un diseño en el cual cada tratamiento aparece una y sólo una vez en cada fila (1a fuente de variación), y una y sólo una vez en cada columna (2a fuente de variación).

Debemos recordar que aunque trabajamos con dos restricciones a la aleatoriedad, seguimos trabajando con solo un factor.

Este diseño es posible únicamente cuando el tamaño de cada una de las restricciones es igual al número de tratamientos.

El tipo de ANOVA a usarse es el de tres vías.

En caso de tener tres restricciones a la aleatoriedad, estaremos frente a un diseño de cuadrado greco-latino.

Experimentos Factoriales.-

Llamamos experimento factorial a aquel en el cual todos los niveles de un factor (variable independiente 1) son comparados con todos los niveles de otro(s) factor(es) (variable independiente 2(n) ).

El modelo matemático usado vendrá dado por:

$$Y_{ijk} = \mu + \tau_{ij} + \rho_k + \varepsilon_{ij}$$

en donde tij es el efecto del iésimo tratamiento del 1er factor con el jesimo tratamiento del 2do factor y rk el efecto de la kesima repetición.

Para 2 factores se puede descomponer tij, dando lo siguiente:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \rho_k + \varepsilon_{ij}$$

En donde ai es el efecto del 1er factor en el iesimo nivel, bj el efecto del 2o factor en el jesimo nivel y (ab)ij el efecto de la interacción de ambos tratamientos en sus respectivos niveles.

La diferencia de un experimento factorial con otro diseñado en bloques, es que en el diseño en bloques nos interesa solamente los efectos de un factor aunque lo separemos en bloques para eliminar las interferencias de otro factor. En un experimento factorial, en cambio, estamos interesados en evaluar el efecto de estos dos o más factores, así como sus posibles interacciones.

Una interacción ocurre cuando un cambio en un factor produce un distinto cambio en nuestra variable a un nivel de otro factor que a otro nivel de este otro factor.

Algunas de las ventajas de los diseños factoriales son:

- Es posible mayor eficiencia que con el diseño de experimentos de un factor.
- Todos los datos son usados para calcular todos los efectos.
- Se adquiere información sobre posibles interacciones entre los dos tratamientos.

Veamos un ejemplo un poco cómico pero ilustrativo para discusión:

Un estadístico borracho desea averiguar cual es la causa de sus chuchaquis, por lo que hace el siguiente experimento:
La primera noche toma whiskey con agua, la segunda ron con agua, la tercera Gin con agua y la cuarta vodka con agua. Como todas las mañanas amanece

mal concluye que el culpable de los malestares fue el agua, ya que este fue el factor común.
Que opina de esta conclusión? Como haría Ud. este experimento?

Determinación del tamaño muestreal.-

Uno de los principales factores que hay que considerar en nuestro diseño experimental es el del tamaño de la muestra.

En general, es recomendable tomar la muestra tan grande como sea posible; sin embargo, podemos calcular matemáticamente el número mínimo necesario con base en lo siguiente:

- La mínima diferencia que deseamos detectar ($\Delta$).
- La variación existente en la población ($\sigma$).
- La máxima probabilidad de error que deseamos tomar ($\alpha$ y $\beta$).

Para estimaciones de parámetros, nuestro tamaño muestreal vendrá dado por:

$$n = [\frac{Z_{(\frac{\alpha}{2})}\sigma}{\Delta}]^2$$

Para pruebas de hipótesis tenemos la fórmula:

$$n = [\frac{(Z_{(\alpha)} + Z_{(\beta)})\sigma}{\Delta}]^2$$

para ensayos de una cola, y la misma usando $Z(\alpha/2)$ y $Z(\beta/2)$ para ensayos de dos colas.

A pesar de esto, los tamaños muestreales son muchas veces escogidos de forma arbitraria, debido a limitaciones de orden económico y práctico.

Datos atípicos.-

Llamamos datos atípicos a aquellos datos que debido a su lejanía de los otros datos de nuestra muestra, podemos considerarlos como no pertenecientes a la población, o como datos que desvían nuestra muestra del verdadero valor de la población.

Es importante tener un criterio objetivo matemático ya que considerar subjetivamente los mismos puede tener serias consecuencias.

Existen varios criterios para determinar que números son atípicos, generalmente en base del tamaño de la muestra y de su varianza. Nosotros estudiaremos el criterio de Chauvenet.

Este criterio considera atípicos a los datos que se alejan de la media más de 1/2n de lo que lo haría una población normal.

Calculamos para cada dato el estadístico:

$$\frac{|x_m - \bar{x}|}{s}$$

y, si el valor del mismo es mayor que el valor correspondiente en la tabla, consideramos al dato atípico y lo eliminamos.

Hay que tener cuidado de no a realizar este procedimiento más de una vez en una muestra, porque corremos el riesgo de que al disminuir nuestra desviación estándar eliminemos datos que no sean atípicos.

Otros criterios para determinar datos atípicos son el test de Dixon y el de Grubb.

**Estadística Predictiva.-**

Una de los principales usos que se trata de dar a la estadística es la predicción de eventos futuros. Exísten muchos métodos de todo tipo que tratan de predecir lo que ocurrirá en condiciones dadas. Todos tienen sus pros y sus contras, no existe un solo método que nos asegure que las predicciones que obtengamos van a ser reales. Especialmete porque las variables que podemos estar considerando fijas y en las cuales basmos nuestras predicciones pueden cambiar. Veremos en esta sección tres métodos que pueden ser útiles en condiciones dadas, se tratan de ajustes de curvas, que ya vimos cuando hablamos de regresiones, arboles de decisión, de los cuales ya vimos la base al hablar de probabilidades, y series de tiempo método que utiliza fundamentos de regresión y otros procedimientos para descomponer en varios coponentes la variación que sufre a lo largo del tiempo un valor. Además veremos las bases de simulación, método que nos permite evaluar propabilidades en condiciones complejas.

<u>Ajustes de Curvas</u> (1 Hora)

Ya vimos el método de los mínimos cuadrados para determinar ajustes de datos a curvas. En esa sección veremos como estimar límites de predicción para las curvas ajustadas mediante dicho método. Esto nos permitirá tener límites probabilísticos para nuestras estimaciones. De esta forma podremos predecir con un intervalo de confianza como se comportará nuestra variable.

Es muy conveniente tener un predicción de un valor fururo de y para un valor dao de x que este dentro del rango de experimentación. Es importante el "dentro del rango de experimentación", ya que la extrapolación es siempre arriesgada y no siempre se mantienen las relaciones fuera de los rangos en que se realizó el experimento.

Veremos 2 casos, el primero es el de estimar el intervalo de confianza para una estimación puntual de y con respecto a x, y el segundo es el de determinar una banda de confianza para la recta de regresión.

<u>Intervalo de Confianza.-</u>
Cuando queremos construir un intervalo de confianza en el cual puede esperarse que una futura observación se encuentre y se desee una probabilidad determinada para ese valor de x podemos considerar los siguientes límites de predicción para y en ese valor de x:

$$Y = a + bx \pm t_{\alpha/2} s_e \sqrt{1 + 1/n + \frac{n(x_i - \bar{x})^2}{S_{xx}}}$$

En donde $t_{a/2}$ es el valor de "t" de Student con $\nu$ = n-2 grados de libertad, y Se la raiz cuadrada del estimador de varinza de y $S^2_e$ :

$$S_e^2 = \frac{SxxSyy - (Sxy)^2}{n(n-2)Sxx}$$

Y Sxx, Sxy y Sxy son:

$$Sxx = n\sum x_i^2 - (\sum x_i)^2$$

$$Syy = n\sum y_i^2 - (\sum y_i)^2$$

$$Sxy = n\sum x_i\, y_y - (\sum x_i)(\sum y_i)$$

Si queremos en cambio determinar los límites de predicción para la media de y en ese valor usamos la siguiente ecuación:

$$Y = a + bx \pm t_{\alpha/2} s_e \sqrt{1/n + \frac{n(x_i - \bar{x})^2}{S_{xx}}}$$

Bandas de predicción:

A veces es conveniente determinar una banda de confianza sobre toda la recta dentro de la cual el 100x$\alpha$% de las veces la recta real de regresión y=$\alpha$+$\beta$x caiga, los valores para esta banda son:

$$Y = a + bx \pm \sqrt{2F_\alpha} S_e \sqrt{1/n + \frac{n(x_i - \bar{x})^2}{S_{xx}}}$$

En donde F$\alpha$ es el valor de F de Fisher con $\nu$1= 2 y $\nu$2=n-2

y esperamos que la media de y para los diferentes x caiga en esa banda el 100x$\alpha$% de las veces.

Series de Tiempo (2 Horas)

Muchos eventos se comportan de forma tal respecto al tiempo que los factores de variación pueden separarse en varios componentes:

- Componente de Tendencia Lineal a largo plazo (por ejemplo crecimientos o disminuciones a largo plazo)
- Componente de un efecto estacional (por efecto de estaciones establecidas que afectan en determinados períodos del año)
- Componente de un efecto cíclico (por efecto de otras variaciones que no son estacionales pero que se repiten de una forma cíclica)
- Componente del error experimental (por efecto del error experimental, las cuales deben de tener una distribución con media 0)
- Otros componentes de error (por efecto de otras variables que no podemos controlar o analizar y que afectan a nuestras estimaciones)

De estos componentes, el último es el único que no podemos analizar, aunque en algunos casos si podamos identificarlo.

En el caso de estar frente a un evento que se comporte de esta manera, los métodos de series de tiempo permiten separar los diferentes componentes de la variación para evaluarlos en forma separada, y una vez hecho esto, calcular la contribución de cada uno de ellos al valor esperado y combinarlos de nuevo para determinar nuestra predicción.
Los métodos de series de tiempo son especialmente útiles en ayudar a analizar el comportamiento de eventos que ocurren con un efecto cíclico o estacional.

Analizaremos 2 métodos de serie de tiempo: el suavizamiento de curvas y la determinación de índices estacionales.

Suavizamiento de curvas:

Al analizar los datos de una serie de tiempo, encontramos que los datos presentan la variación de tal forma que es dificil a simple vista separarlas. El suavizamiento de curva es in método que permite eliminar variaciones a corto plazo para evaluar el comportamiento a mayor plazo de la curva.
Hay varias formas de suavizamiento. Veremos 2, el método de pormedios móviles y el de suavizamiento exponencial.

El método de promedios moviles consiste en definir un período de tiempo tal que permita eliminar las variaciones externas a la tendencia. De esta forma si queremos por ejemplo evaluar un fenómeno con variaciones estacionaes dentro del año, usaríamos promedios móviles con una amplitud de 12 meses. El promedio móvil calcula los promedios de los primeros 12 meses, luego calcula del mes 2 al 13, 3 al 14 y asi sigue. La curva resultante será mas “suave” o con menos picos.

El suavizado exponencial asigna pesos a las variaciones actuales y las pasadas, y calcula un promedio ponderado en base a esos pesos, logrando de esta manera una curva que será mas suave y mas lenta a cambios o menos suave y mas sensible a cambios, dependiendo del valor que se le de a los pesos actuales y pasados.

Al analizar series de tiempo, una curva suavizada puede ser analizada por regresión para determinar su tendencia.

Estos métodos no solo sirve para analizar series de tiempo, sino también para suavizar curvas en general donde hay variaciones que queremos pasar por alto. Un uso práctico es en muestreos de pesos en camaroneras en donde los pesos promedios suben y bajan con las fases del aguaje por cuestiones de muestreo. En este caso se puede tener una curva suavizada que nos permita saber cual es la tendencia real del crecimiento.

Determinación de la estacionalidad
Un método de determinación de la estacionalidad es el de índices estacionales. En este método quitamos la influencia de otros factores de variación a nuestra información, quedandonos únicamente con los efectos de la estación y del error.
Agrupamos los datos en las estaciones y calculamos la median de los mismos, determinando luego un índice de estacionalidad que será el efecto de la estacionalidad medido como un porcentaje del valor esperado para dicho momento.

Toma de Decisiones.-
Consideramos un **problema** como una situación que nos impide lograr nuestros objetivos y que puede ser solucionada por nosotros.
Un modelo de decisión debe de considerarse meramente como un vehículo para “resumir” un problema de decisión, en forma tal que haga posible la identificación y evaluación sistemática de todas las alternativas de decisión del problema.

Después se llega a una decisión seleccionando la laternativa que se juzgue sea la "mejor " entre todas las opciones disponibles. Un modelo de toma de decisiones es el siguiente:

1. Identificar y describir clarament el **problema.**
2. Identificar las **alternativas** de decisión y las **restricciones** que se aplican.
3. Diseñar un **criterio** para evaluar el "valor" de cada alternativa.
4. Utilizar el criterio generado como base para seleccionar la mejor de las alternativas disponibles.

Hay que tener siempre en cuenta que cualquier modelo de toma de decisiones, sin importar su refinamiento y exactitud, pueden resultar poco prácticos si no están respaldados por datos confiables. Si se distorsionan las estimaciones, la solución que se obtenga, por mas que sea óptima desde el punto de vista matemática, puede tener poca o ninguna utilidad desde el punto de vista real.

Existen problemas que pueden modelarse matematicamente y otros que no. Ambos sin embargo necesitan de una decisión. Para los problemas que son modelables matematicamente, este modelo no representa la solución, sino que nos sirve de pauta para que nosostros podamos tomar una decisión lógica al respecto.

Existen diversos tipos de modelos para toma de decisiones. Hay modelos que utilizan información perfecta y permiten tomar decisiones en condiciones de **certeza**. Por ejemplo si sabemos los costos de transportar una pesca por camión y por lancha, podemos decidir por el que mejor satifaga nuestro criterio de evaluación, por ejemplo el que tenga menor costo.
Sin embargo, en la mayoría de los casos, poseemos información incompleta o parcial. Esto nos lleva a Otros dos tipos de decisión: Decisión bajo condiciones de **reisgo**, es cuando conocemos los posibles valores que puede tomar la medidad y la probabilidad de ocurrencia de la misma (distribución de probabilidad). Por ejemplo si nuestro camino de entrada no es lastrado y el noticiero nos indica que hay una probabilidad de 30% de lluvia, podemos tomar una decisión bajo riesgo, esto es una decisión en la cual conocemos la probabilidad que nuestra decisión sea la equivocada. En caso que desconoscamos la probabilidad de ocurrencia de lluvia, pero sabemos que es posible, estaremos bajo una decisión bajo **inceridumbre**.
Podemos considerar las decisiones bajo certeza y bajo incetidumbre como los extremos en cuanto a información y las decisiones bajo riesgo como un punto intermedio.

La mayor parte de los modelos de toma de decisión (x.ej pruebas de hipótesis) que hemos repasado son básicamente modelos de toma de decisión bajo riesgo. En esta sección veremos dos métodos más: Cálculo de valor esperado (que ya vimos ligeramente) y simulaciones.

Cálculo de Probabilidades y Valor Esperado, Arbol de decisiones

Ya vimos como se puede calcular el valor esperado de un suceso, pero los casos que tratamos se referían a alternativas de “una etapa”, en el sentido que ninguna decisión a futuro dependerá de la decisión que tomemos ahora. En esta secció consideraremos un proceso de decisión de “múltiples etapas”, en el cual se toman decisiones dependientes unas de otras. Para facilitar los cálculos y el entendimieto de este modelo se suele usar una herramienta gráfica conocida com “**arbol de decisión**”. Esta representación facilita el proceso de toma d e decisiones.

Un arbol de decision es una representación gráfica del problema mediante lineas y nodos. Existen 2 tipos de nodos:

1. Los Nodos de **Decisión** (• ), que representan alternativas sobre las cuales debemos de tomar alguna decisión. Aqui cada alternativa lleva asociada un costo $C_i$.
2. Los Nodos **Probabilísticos** (o), que representan los distintos eventos que son probables que ocurran en dicho evento. Aqui cada evento lleva asociado una probabilidad $P_i$, donde la suma de las probabilidades de todos los eventos ebe de ser igual a 1. A cada Nodo probabilistico se lo califica con su valor esperado $\Sigma C_i P_i$.

Ademas, el arbol de decisión tiene lineas que representan las alternativas o eventos.

El proceso de este método empieza con la construcción del arbol, partiendo de un nodo de decisión actual, del cual parten las distintas alternativas (lineas) entre las cuales debemos de decidir en este momento. De estas alternativas, pueden nacer nodos de decisiones sobre alternativas futuras o de posibles eventos futuros.

Una vez que se representaron todas la alternativas y evntos, empezamos colocando los costos de las últimas alternativas, y del final al comienzo vamos colocando las probabilidades en cada evento, y calculando los valores esperados en cada nodo probabilístico, en caso de encontrar un nodo de decisión, el valor de dicho nodo va a ser la decisión mas conveniente (la de menor costo). Esto lo realizamos hasta llegar a las alternativas del nodo de decisión inicial, en donde podemos tomar la que mas nos convenga.

Veamos un ejemplo:

Un inversionista tiene en este momento $5,000,000 disponibles en este momento, y está considerando la opción de construir en este momento una planta de proceso de una nueva especie de cultivo. En caso de decidirse por este proyecto, tiene la alternativa de construir una planta de tamaño piloto que se pueda ampliar después o de tamaño completo. La decisión de cual planta construir dependerá de las futuras demandas y ofertas de producto a procesar. La construcción de una planta de tamaño grande se justificaría económicamente si el volumen de producto en el futuro es grande. En caso contrario sería conveniente

construir primero una planta pequeña y, luego de 2 años, si las condiciones son propicias, ampliarse.
El problema de múltiples etapas se presenta aquí porque si el inversionista decide construir ahora la planta piloto, en dos años debe decidirse si amplia o no. Dicho de otra manera, el proceso de decisión implica tres etapa:

1. Invertir el dinero en la planta o dejarlo en el banco.
2. En caso de decidirse por construir la planta ¿De que tamaño construir la planta?
3. En caso de construir la planta piloto ¿Deberá de ampliarse después de 2 años?

En el gráfico siguiente se puede ver el arbol de decisión. Empezando con el primer nodo • , el inversionista deberá de decidir si invierte en la planta o deja su dinero en el banco ganando 10% de interes anual. El nodo 2 • , es otro nodo de decisión, en el cual el inversionista deberá de decidir si construye la planta piloto o la grande. El nodo 3 o, es un evento probabilistico, del cual emanan dos ramas que representan los volumenes altos y bajos respectivamente. El nodo 4 • , representa otra decisón, la cual solo deberá de hacerse en el caso de que se decidió por la planta piloto y que los volumenes a procesar sean altos. Una vez mas, en caso de ampliar o no la planta, existe la posibilidad que después de dos años los volumenes se mantengan altos o de que declinen, lo cual esta representado por el nodo 5 o y las ramas que emanan de el (alto o bajo).

Los datos del arbol de decisión deben de incluir:

1. Las probabilidades asociadas con las ramas que emanan de los nodos probabilisticos.
2. Los ingresos y costos asociados con las distintas alternativas del problema.

Supongamos que el inversionista está interesado en estudiar el problema en un período de 10 años. Un estudio de mercado señala que las probabilidades de tener volúmenes de producción altos y bajos respectivamente son de 0.75 y 0.25 respectivamente, pero en caso de que los volumenes sean altos durante los primeros 2 años, la probabilidad de que se mantengan altos sube a 0.90, con una probabilidad de que bajen de 0.10. El estudio de mercado indica también de que en caso de que los volumenes no despeguen en los dos primeros años, la probabilidad de que mejoren después es despreciable ($p \approx 0$). La construcción de una planta grande costará \$5,000,000 y de la pequeña \$1,000,000. Se calcula que la expansión de la planta de aquí a 2 años costará \$4,000,000, aunque la planta resultante de esta expansión no tendría la misma capacidad de procesamiento que la planta construida inicialmente grande. Las estimaciones de ingresos anuales son las siguientes:

1. La planta grande y un volumen alto producirá \$1,000,000 anuales.
2. La planta grande y volúmenes bajos producirá \$300,000 anuales.
3. La planta pequeña y un volúmen alto producirá \$250,000 anuales.
4. La planta pequeña y un volúmen bajo producirá \$200,000 anuales.

5. La planta pequeña ampliada con volumenes bajos producirá $200,000 anuales.
6. La planta pequeña ampliada con volumen alto producirá $900,000 anuales
7. El dinero en el banco producirá 10% de interes simple anual, retirando los intereses anualmente.
8. No se considera inflación, y ell valor de recuperación de los activos al final del decimo año se considera del 100%.

Estos datos y las probabilidades asociadas se presentan en el gráfico respectivo.

Los cálculos son los siguientes:

- El dinero en el banco producirá :

+$5,000,000 * 0.10 = $500,000 anuales * 10 años = **$5,000,000**

- La constucción de la planta pequeña costará:

**-$1,000,000**

- La construció de la planta grande costará:

**-$5,000,000**

- Los ingresos que se obtendrán en caso de que se opte por la planta pequeña y la demanda sea baja durante los 10 años será de:

+$200,000 * 10 = +$2,000,000
Mas los intereses del dinero que no invertimos, y dejamos en el banco ganado por diez años :
+$4,000,000 * 0.10 = 400,000 anuales * 10 años = +$4,000,000
lo que nos da un total de:
+$2,000,000 + +$4,000,000 = **+$6,000,000**

- Los ingresos que se obtendrán en caso de que se opte por la planta grande y que los volumenes sean bajos por 10 años serán de:

+$300,000 anuales * 10 años = **+$3,000,000**

- Los ingresos que obtendremos en la planta pequeña por 2 años con volumenes altos son de:

+$250,000 anuales * 2 años = +$500,000
Mas los intereses :
+$1,000,000 * 0.10 = 100,000 anuales * 2 años = +$200,000
lo que nos da un total de :
+$500,000 + $200,000 = **+$700,000**

- Los ingresos que obtendremos por 2 años en la planta grande con volumenes altos será de :

$1,000,000 anuales * 2 años = **+$2,000,000**

- Los ingresos a obtener por 8 años en la planta grande con volumenes que se mantengan altos son de:

+$1,000,000 anuales * 8 años = **+$8,000,000**

- Los ingresos que se obtendrán en los 8 años de la planta grande con volúmenes que disminuyan serán de :

+$300,000 anuales * 8 años = **+$2,400,000**

- Los ingresos que se obtendrían durante 8 años en caso de ampliar la planta y mantener los volúmenes altos sería de:

+$900,000 anaules * 8 años = **+$7,200,000**

- Los ingresos que se obtendrían durante 8 años en caso de ampliar la planta y que los volúmenes decaigan sería de :

+$200,000 anuales * 8 años = **+$1,600,000**

- Los ingresos que obtendríamos en caso de que no se amplíe la planta y que los volumenes se mantengan altos por los 8 años restantes serían de:

+$250,000 anuales * 8 años = +$2,000,000

Mas los ingresos por intereses:
+$4,000,000 * 0.1 = +$400,000 anuales * 8 años = +$3'200,000

Lo que da un total de :
+$2,000,000 + $3,200,000 = **+$5,200,000**

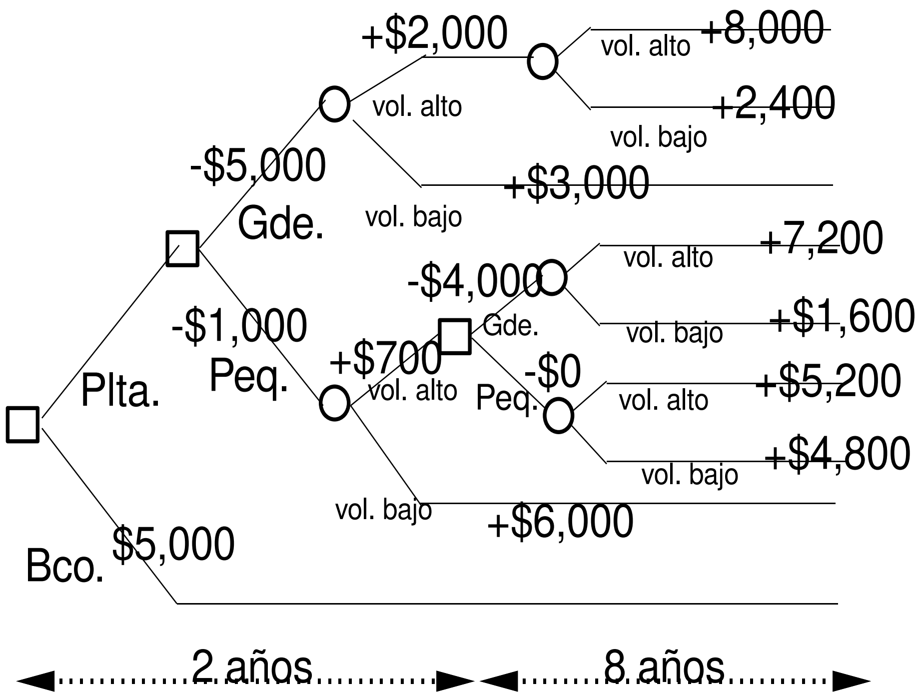
- Los ingresos que obtendríamos en caso de que no se amplíe la planta y que los volumenes decaiganse por los 8 años restantes serían de:

+$200,000 anuales * 8 años = +$1,600,000

Mas los ingresos por intereses:
+$4,000,000 * 0.1 = +$400,000 anuales * 8 años = +$3'200,000

Lo que da un total de :
+$1,600,000 + $3,200,000 = **+$4,800,000**

Estos valores deberán de ser coloados en el arbol de decisión:




Una vez hecho esto podemos calcular los valores esperados hasta los primeros nodos de decisión:

- Para la construcción de la planta grande será:

Primero calculamos la esperanza en volumen alto:
+$8,000,000 * 0.9 + $2,400,000 * 0.10 + $2,000,000 = <u>+$9’440,000</u>
 y calcuamos la esperanza por ingreso:
+$9,440,000 * 0.75 + +$3,000,000 * 0.25 = +$7,830,000
Luego quitamos el valor de la inversión:
+$7,830,000 - $5,000,000 = **+$2,830,000**

- Para la ampliación de la planta pequeña será:

+$7,200,000 * 0.90 + 1’600,000 * 0.10 = +$6’640,000
Menos el costo de la ampliación:
+$6,640,000 - $4,000,000 = **+$2,640,000**

- Para la no ampliación de la planta pequeña será :

+$5,200,000 * .90 + $4,800,000 * 0.10 = **+$5’160,000**

Aqui llegamos al último nodo de decisión, y escogemos la mejor alternativa, en donde el nodo tomará este valor, osea **+$5,160,000**

Y calculamos el valor del siguiente evento:

- Para la construcción de la planta pequeña:

(+$5,160,000 + $750,000)* 0.75 + $6,000,000 * 0.25 = <u>+$5’932,500</u>
menos el costo de la construcción:
+$5,932,500 - $1,000,000 = **+$4,932,500**

Aqui llegamos al segundo nodo de decisión, el de decidir el tamaño de la planta, y como la eperanza al construir la planta pequeña (+$4,932,500) es mayor que la de construir la planta grande (+$2,830,000), decidimos por la planta pequeña y asignamos este valor al evento de construir.

Llegamos entonces al primer nodo de decisión, en el cual debemos de decidir entre construir una planta con una esperanza de ingreso de +$4,932,500 o de dejar la plata en el banco con una esperanza de ingreso de +$5,000,000 (nótese que como se considera que todas las opciones presentan la misma oportunidad de recuperar los 5 millones de inversión inicial estos no se los está considerando). Como la opción de dejar la plata en el banco tiene una mejor esperanza, esta sería la opción sugerida por el modelo a escoger.

Hay que notar que este ejemplo no es totalmente correcto desde el punto financiero, ya que en realidad deberíamos de considerar los flujos de ingresos y egresos en el tiempo (considerando el costo de oportunidad), pero sirve bien como un ejemplo desde el punto de vista probabilístico.

<u>Simulación</u> (1 hora)

La simulación es un método que nos permite evaluar como se comportará un sistema complejo bajo ciertas suposiciones. Es muy util por cuanto aprovecha la capacidad de los sistemas informáticos actuales para realizar gran cantidad de cálculos complejos rápidamente.
Un modelo de simulación busca imitar el comportamiento de istema que se investiga estudiando las interacciones entre sus componentes. El “output” o salida de un modelo de simulación se presenta normalmente en términos de medidas seleccionadas que reflejan el desempeño del sistema.
La simuación debe de tratarse como un experimento estadístico. A diferencia de los modelos matemáticos que ya revisamos, en donde el “output” del modelo representa un comportamiento estable a largo plazo, los resultados que se obtienen al ejecutar un modelo de simulación son observaciones que están sujetas al error experimental. Esto significa que cualquier inferencia relativa al desempeño del sistema simulado debe de estar sujeta a todas las pruebas adecuadas de análisis estadístico.
Un experimento de simulación difiere de un experimento normal, en que se puede realizar totalmente en la computadora. Al expresar las interaciones entre los

componentes del sistema como relaciones matemáticas, podemos recopilar la información necesaria casi en la misma forma como si estuvieramos observándolo en el sistema real (sujeto desde luego a las simplificaciones integradas en el modelo). Por lo tanto, la naturaleza de la simulación permite mayor flexibilidad en la representación de sistemas complejos que son dificiles de analizar mediante modelos matemáticos estandar. Si embargo, debemos de tener en cuenta que aunque la simulación es una técnica flexible, la elaboración de un modelo de simulación puede demorar mucho y ser muy costosa, en especial cuando se trata de optimizar el modelo. Además requiere de conocimientos del comportamiento de las principales variables del sistmema.
No es objetivo de este seminario dar un detalle de la forma de realizar simulaciones, pero el siguiente ejemplo dará una idea de sus posibles usos.

## BIBLIOGRAFIA


- **Capa H. (1991)**. *Estadística Básica*. Cimacyt.
- **Hicks C. (1982)**. *Fundamental concepts in the design of experiments*.CBS College Publishing.
- **Leedy P. (1974)**. *Practical Research: Planning and design*. Macmillan.
- **Marcillo F. (1992).** *Manual Práctico de Estadística Básica y Diseño Experimental Aplicados a la Acuacultura*. Centro de Educación Contínua, ESPOL
- **Miller I., Freund J. (1986).** *Probabilidad y estadística para ingenieros*. Prentice - Hall.
- **Möller F. (1979).** *Manual of methods in aquatic enviroment research. Part 5.- Statistical tests.* FAO Fisheries technical paper No. 182.
- **Scheaffer R., Mendenhall W., Ott L. (1987).** *Elementos de Muestreo*. Grupo Editorial Iberoamérica.
- **Steel R., Torrie J. (1985)**. *Bioestadistica: Principios yprocedimientos.* McGraw - Hill.
- **Taha H. (1991).** *Investigación de Operaciones*. Alfaomega