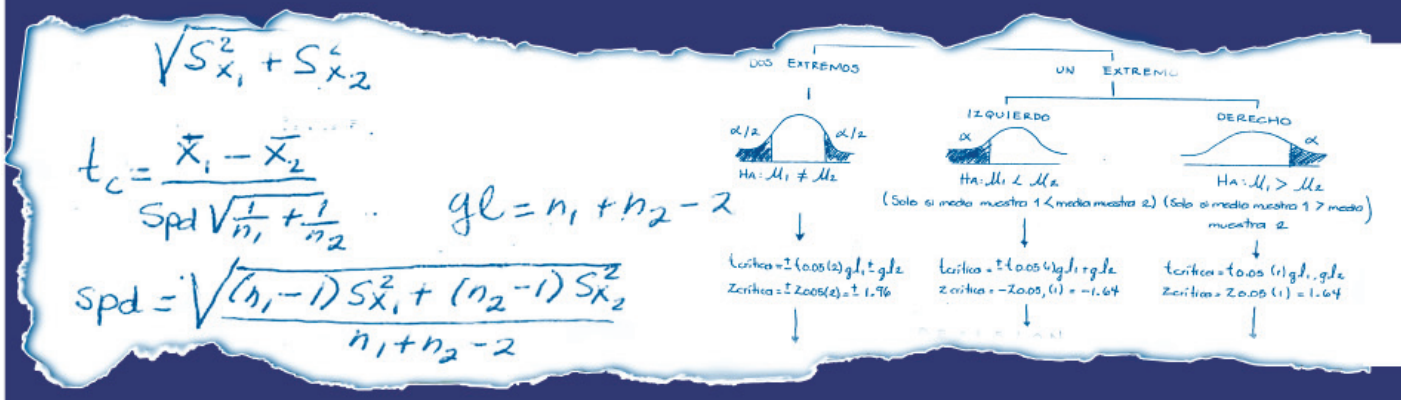
Estadística descriptiva I

Probabilidad y pruebas de hipótesis



Compiladores

Domingo Flores Hernández
Julia Ramos MIranda
Atahualpa Sosa López

UNIVERSIDAD AUTÓNOMA DE CAMPECHE

Estadística descriptiva probabilidad y pruebas de hipótesis

I

Compiladores:

Domingo Flores Hernández

Julia Ramos Miranda

Atahualpa Sosa López

Estadística I

Flores Hernández, D., J. Ramos Miranda y A. Sosa López (Compliladores), 2007. Estadística Descriptiva, Probabilidad y Pruebas de Hipótesis I. Universersidad Autónoma de Campeche. Facultad de Ciencias Químico Biologícas. ISBN xx-xxx-xxx.150 p.

Facultad de Ciencias Químico Biológicas - Centro EPOMEX

ISBN

Contenido

PRESENTACIÓN

La presente obra fue realizada con el objeto de brindar mediante explicaciones simples y con varios ejemplos biológicos, las herramientas necesarias para que los alumnos de la carrera de Biología se inserten en el campo de la estadística. En efecto una de las principales dificultades con las que se encuentran los alumnos al consultar los diferentes libros de estadística es la gama de definiciones sobre los conceptos y las diferentes notaciones para los mismos, esto dificulta la rápida comprensión de la estadística y por lo tanto su aplicación.

Para su práctico aprovechamiento, esta obra esta diseñada conforme al plan de estudios de la materia y compila aspectos de varios compendios de estadística aplicada a la biología, y consultas vía Internet. La obra incluye las generalidades de la estadística, diseños de plan experimental y la estadística descriptiva para continuar con las distribuciones de probabilidad (Binomial, Poisson, t- Student, χ^2 y Fisher) e insertar los conceptos de estadística inferencial, comparación de muestras paramétricas y no paramétricas y finaliza con el cálculos de correlación y regresión lineal simple.

La compilación de esta información ha sido en muchos casos traducida al español para la rápida comprensión de los alumnos. Esperamos que la selección de textos aquí presentados sea valiosa y útil para comprender y aplicar la estadística principalmente en el área de la Biología.

Los Compliladores

1 Introducción

1.1. Breve Historia de la Estadística

Literatura sugerida:

Scherrer B., 1984, Biostatistique. Gaëtan Morin Editor. Montreal, Paris, Casa Blanca. 850 p (Pág 1).

Zar. J. H., 1999. Bostatistical Analysis. 4 edición. Prentice Hall. Estados Unidos. 663 p. (Pag. 1)

Estadística proviene del latín "Status" y "Statisticus" (relativo al Estado), refiriéndose a Estados Políticos. La necesidad de colectar datos estadísticos sobre la población y sus condiciones materiales se remonta a la antigüedad. De esta manera se realizaron los inventarios de población y de productos agrícolas. Se citan frecuentemente los ejemplos del emperador chino Yao Qui, en 2238 A.C., quien organizaba el reconocimiento de los productos agrícolas: el Faraón Amasis, quien dictaba la pena de muerte contra los que se rehusaban a declarar su nombre, profesión y medios de subsistencia. En Grecia y la antigua Roma, se habla de censos en las fechas den nacimiento de "Cristo", navidad. Cuando el emperador Augusto ordena la reentrada de sus guerreros, cada soldado debería regresar a su lugar de origen para ser registrado. En los siglos XIII y XIV el comercio de Venecia, utilizó la estadística para contabilizar el paso de los productos. Posteriormente los hermanos Elzevir, de los Paises Bajos, publicaron a principios del Siglo XVII una enciclopedia en sesenta volúmenes sobre la economía y el comercio de los estados.

Los primeros conceptos de estadística aparecen en Alemania en el VII siglo, creando la palabra "Statistik" y desarrollan algunas nociones para fines académicos. En Inglaterra, J. Graunt, W. Petty y E. Halley de la escuela de política aritmética, se enfocaron sobre todo al aspecto matemático de los seguros (tablas de mortalidad) y ponen en evidencia ciertas estadísticas que van más allá de una simple descripción de datos. Finalmente en Francia, Colbert y Vauban, ejecutan numerosos inventarios y reconocimiento de la población y sus recursos. En la actualidad el uso de la estadística es cotidiano y sin pensarlo siempre se hace uso de ella. Esto nos sirve en realidad para ordenar y analizar nuestro entorno.

1.2 Estadística y su Importancia en la Investigación Científica

1.2.1 Definición

La estadística es un **conjunto de teorías y métodos científicos que han sido desarrollados para tratar la recolección, el análisis y la descripción de datos con el objeto de extraer conclusiones útiles para la solución de algún problema en particular de algún universo colectivo**. Comúnmente

Literatura sugerida:

Daniel.W. W. 1982, Biostadística. Limusa. Mexico. 485 p (pág. 1)

Sokal R. R. y Rohlf F. J. 2000. Biometry. 3a edición. Ed. Ferman. Estados Unidos. 887 p (pág. 2).

Zar. J. H., 1999. Bostatistical Analysis. 4 edición. Prentice Hall. Estados Unidos. 663 p. (Pag. 1).

http://euler.ciens.ucv.ve/pregrado/estadistica/archivos/guias-teo/guia1.pdf#search='estad%C3%ADstica%20descriptiva'

la estadística apoya al investigador a inferir sobre los parámetros de la población a partir de estadísticos muestrales.

La estadística se subdivide en dos áreas denominadas:

a) **Estadística descriptiva** y,

b) **Estadística inferencial**.

La estadística descriptiva llamada también deductiva, permite obtener de un conjunto de datos, conclusiones que no sobrepasen la información que proporcionan los mismos datos; su estudio incluye las técnicas de colecta, ordenamiento, análisis e interpretación de datos. Es decir, organiza y resume observaciones que sean fáciles a comprender.

La estadística inferencial o inductiva, es el conjunto de técnicas utilizadas para obtener conclusiones que rebasan los límites de la información aportada por los datos; así mismo a través del uso de dichas técnicas se busca obtener información de un universo colectivo a partir de datos tomados de él.

2 Estadística Descriptiva

2.1 La estadística en el Enfoque Metodológico de la Investigación

Literatura sugerida:

Scherrer B., 1984, Biostatistique. Gaëtan Morin Editor. Montreal, Paris, Casa Blanca. 850 p (Pág 29-98).

Un plan de investigación ideal, debería proporcionar resultados **no triviales** para que tengan alguna utilidad, **no equívocos**, para no dar lugar a interpretaciones diversas, y **ser generalizables**. Para esto, se requiere una planificación detallada, la cual puede exigir más tiempo y esfuerzo que el desarrollo mismo de la investigación. El investigador que por intuición repentina colecta datos, puede ahorrarse un tiempo considerable, pero corre el riesgo de obtener resultados evidentes desde el inicio o difícilmente interpretables.

No hay una única forma de hacer un plan de investigación. El procedimiento aquí presentado; propuesto por Scherrer (1984) es sólo una guía. El número, la secuencia, la importancia y la naturaleza de las etapas varían de un investigador a otro, según su disciplina, su formación. El proceso de la figura 1, divide arbitrariamente las etapas para facilitar la presentación para comprender mejor lo intrincado de los métodos estadísticos y el conjunto del proceso de investigación.

2.1.1 Definición del Problema

Toda investigación surge de un problema: es decir de una serie de dificultades resultando de una desviación de la situación actual respecto de la deseada o esperada.

El análisis preliminar de un problema, revela casi siempre una serie de otros problemas unidos o relacionados unos con otros, los cuales en conjunto reciben el nombre de **problemática**.

En esta etapa el investigador se esforzará en formular explícitamente la problemática, con el fin de identificar los problemas o procesos relacionados. Esta formulación orienta en sí el desarrollo de la investigación, ya que según se haya formulado, se condiciona al menos parcialmente la respuesta.

Así en ecología por ejemplo, el problema de muestreo de comunidades vegetales o animales tiene dos grandes obstáculos:

1. Obtener una muestra representativa de la población biológica.

2. Elaborar un plan de muestreo y un dispositivo de colecta que proporcionen una estimación pre-

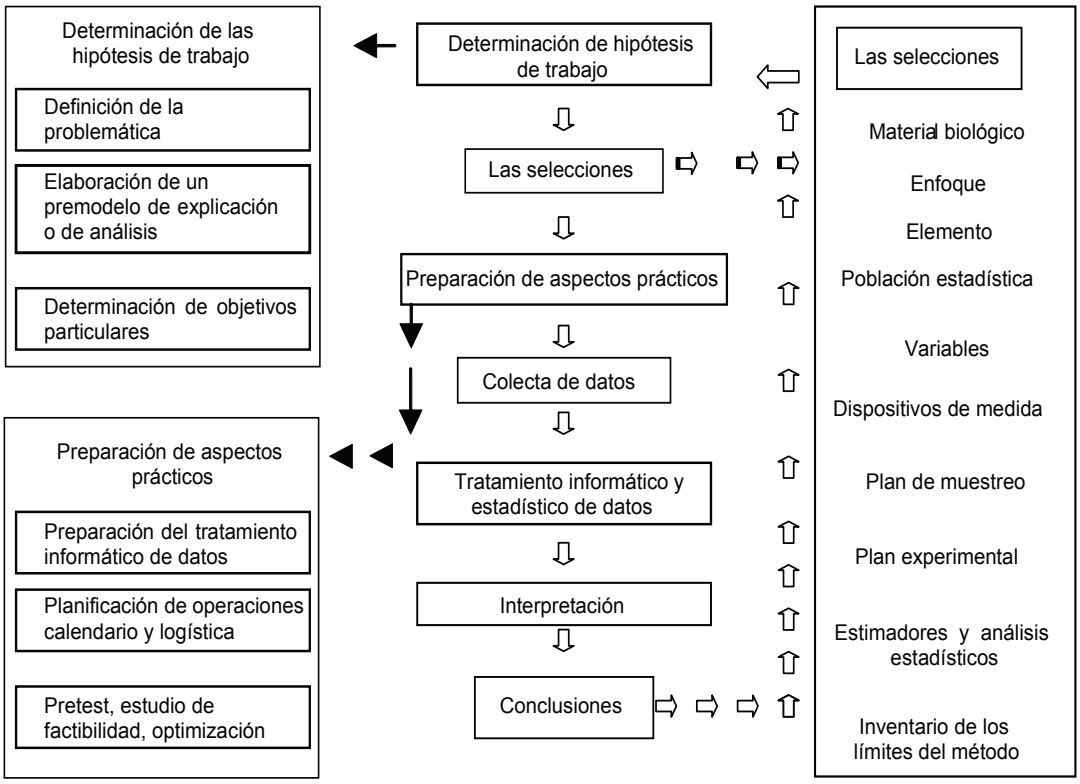


Figura 1. Principales etapas del proceso metodológico (tomado de Scherrer (1984), pág 30).

cisa y no sesgada de la abundancia de cada una de las especies que habitan un área dada en un momento determinado.

En general abordar una problemática completa es muy complejo, esta complejidad incluye al muestreo también. En la práctica es frecuentemente necesario limitarse a uno o pocos aspectos de la problemática general.

2.1.2 Examen del Estado de Conocimientos del Problema

Después de haberse concentrado en un problema específico, es necesario hacer una revisión de la literatura para hacer un balance de conocimiento del problema. Por ejemplo, para las "lluvias ácidas" un estado del conocimiento del problema se establece como sigue:

1. La Comisión Europea de pesca de aguas interiores mostró que un pH inferior a 4.5 es letal para la mayor parte de especies de peces; además, que valores entre 4.5 y 5.0 afectan probablemente la reproducción de los salmónidos.

2. Mount (1973) demostró que la producción y la tasa de eclosión de huevos se reduce cuando el pez *Pimephales promelas* (la cabeza de bola) es expuesto a un pH >5.0 en aguas relativamente duras (200 mg/l de $CaCO_3$).

3. Johansson *et al.* (1973) observaron la reducción de la tasa de eclosión del pez *Brachidanio rerio*, etc.

2.1.3 Elaboración de un Modelo Conceptual de Explicación o Análisis

El modelo conceptual busca proveer una explicación de la problemática. Este integra elementos de literatura, tiene un carácter especulativo o hipotético, ya que generalmente varios supuestos no han sido aún verificados.

La utilidad de este modelo es de proporcionar un marco de reflexión y un camino lógico para proponer una hipótesis de trabajo. Esta hipótesis podrá ser confirmada, redefinida, o refutada para continuar con otras rutas de investigación en el marco de la misma problemática. Este **modelo** o **paradigma** (Kuhn, 1970) se trata de un conjunto *a priori,* más o menos concientes que constituyen el fondo de toda investigación científica.

Después de una serie de investigaciones aceptando el mismo paradigma, aparecen generalmente, un cierto número de **anomalías** e incoherencias en la explicación de los procesos o fenómenos. De estas anomalías, nuevos paradigmas con nuevos *a priori* se desarrollan y así sucesivamente.

> **Ejemplo:** al inicio de los años 80, la investigación agronómica se concentraba en dos tipos de agricultura, una convencional y química, sostenida por la ciencia normal; la otra biológica, marginal. Entre los supuestos de la escuela convencional, figuran los resultados de Liebig (1840), que preconiza una alimentación de la planta con fertilizantes solubles directamente asimilables por la raíces. Además entre los objetivos compartidos de esta escuela, está la maximización del rendimiento agrícola (producción por unidad de superficie). Actualmente este enfoque presenta una anomalía, el empobrecimiento o aún la supresión progresiva del suelo, como es el caso del medio oeste americano y de las praderas canadienses. En contraparte, la escuela biológica tiene como *a priori* el alimentar el suelo y no directamente la planta, claro esto en detrimento del rendimiento.

El **modelo conceptual** (figura 2) puede ser profundizado sobre el plan teórico para construir un **modelo matemático** *a priori*. Estos pueden ser entonces **modelos analíticos**, donde el sistema es definido por ecuaciones resolubles analíticamente. **Modelos de simulación**, donde el sistema es demasiado complejo para ser resuelto de manera analítica y entonces se obtiene una solución numérica. Según otra clasificación, el modelo puede ser **determinista**, si no tiene en cuenta las incertidumbres de los datos; o **estocástico** si se hace intervenir el aspecto aleatorio (probable) de los fenómenos.

2.1.4 Determinación de Objetivos Particulares

En la mayoría de investigaciones los objetivos consisten en verificar las hipótesis de trabajo emitidas en el modelo conceptual. En ciertas investigaciones relacionadas con el manejo de recursos del ambiente, o al establecimiento de estimaciones determinadas, los objetivos del trabajo no contienen *per se* una hipótesis de trabajo (Scherrer, 2001).

Las hipótesis y objetivos de trabajo serán operantes en la medida que posean 7 atributos principales:

1. **Explícitos**, que sean accesibles a todos los que intervienen en el trabajo.
2. **Específicos**, para dar una respuesta única: positiva, negativa o cuantificada.

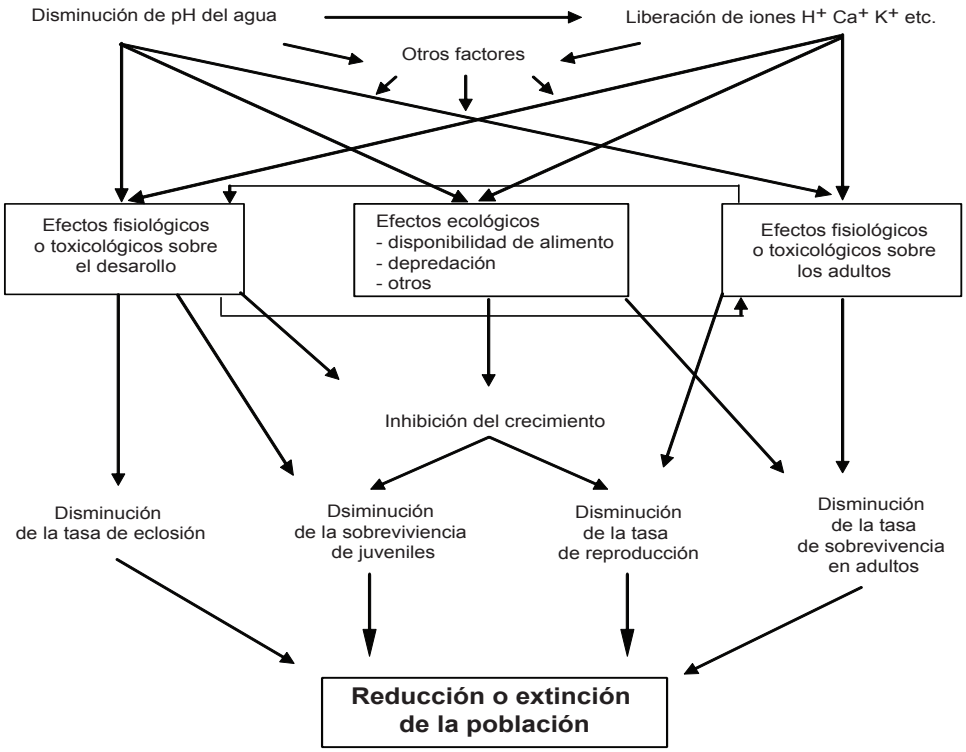


Figura 2. Modelo conceptual de análisis (lluvias ácidas), (tomado de Scherrer (1984), pág 35).

3. **Conformes** a la problemática general del modelo para evitar dispersar esfuerzos en dominios que no contribuyen a la elucidación de la problemática.

4. **Colectivamente exhaustivos y coherentes**, para facilitar la interpretación del conjunto de resultados y obtener una explicación única.

5. **Compatibles**, para que el alcance de un objetivo, no impida la realización del otro.

6. **Realistas** para permitir una adecuación entre los recursos disponibles y las metas establecidas.

7. **Jerarquizados** para permitir obtener un orden de prioridad en su realización.

En el ejemplo del efecto de las lluvias ácidas sobre poblaciones de peces, las preguntas buscadas podrían ser formuladas como sigue:

1. ¿A partir de que pH los ovocitos de hembras adultas no expuestas a estrés por aguas ácidas, tienen una tasa de eclosión menor?

2. ¿A partir de que pH los ovocitos de hembras adultas no expuestas a estrés por aguas ácidas, no eclosionan?

3. ¿A partir de que pH los ovocitos de hembras adultas expuestas a estrés moderado por aguas ácidas, tienen una tasa de eclosión menor?

4. Etc.

2.1.5 Selecciones a Realizar

La elaboración de un enfoque metodológico científico, implica una serie de elecciones y decisiones que tienen todas repercusiones unas sobre otras, y que se referirán en los párrafos siguientes.

2.1.5.1 Selección del Material Biológico

La complejidad depende del tipo de investigación; seleccionar al azar una muestra de organismos cultivados para una serie de experimentos en laboratorio, es una cosa fácil, pero en cambio, es difícil seleccionar organismos silvestres.

En el caso de las lluvias ácidas mencionado antes, el material biológico deberá, en la medida de lo posible, presentar interés económico, ser frecuente en los lagos vulnerables a las lluvias ácidas, disponible sin dificultad mayor y barato. Además de talla reducida para evitar el empleo de acuarios voluminosos, capaz de vivir en acuario sin instalaciones costosas, tener huevos fácilmente recuperables, estar más o menos en extinción o en fuerte disminución. En este caso la mejor opción fue la trucha (*Salvelinus fontinalis*).

2.1.5.2 Selección del Enfoque

Dos enfoques fundamentales están a disposición del investigador: el experimental y el descriptivo; Sin embargo existe también el enfoque cuasi-experimental que será tratado en la sección 2.1.6.

a) Enfoque experimental

Es el método por excelencia para identificar relaciones de causa efecto. Consiste en observar bajo condiciones controladas por el investigador, el efecto de la variación de una o varias variables del fenómeno estudiado. Para su realización, las unidades de muestreo seleccionadas aleatoriamente, son colocadas en un ambiente controlado libres de toda influencia exógena, o al menos que las mantiene constantes (variables que no son estudiadas). Por modificación más o menos progresiva del valor de las variables o factores analizados, se observa la variación de la variable estudiada.

La comparación de resultados obtenidos sobre una muestra **con tratamiento**, es decir sometida a variaciones, con los resultados de muestras aleatorias **sin tratamiento** o testigos, permite despejar con una cierta seguridad las relaciones de causalidad. El mayor **inconveniente** de este enfoque, es que no es fácil extrapolar los resultados de un ambiente artificial al ambiente natural, ya que el comportamiento puede ser diferente.

b) Enfoque descriptivo

Consiste en obtener una imagen tan precisa y fiel como sea posible de un fenómeno particular. Es el método por excelencia para realizar un balance de una situación dada, definir el estado de un sistema o determinar las características estructurales, dinámicas u otras de una población. La red de correlaciones que unen las variables observadas no tienen una connotación de causa efecto, sino más bien de covariación o de correspondencia en sus fluctuaciones.

La elección del enfoque depende esencialmente de tres factores:

1º. De los objetivos buscados, ya que el enfoque descriptivo proporciona la imagen de una cierta realidad, mientras que el enfoque experimental indica una relación de causalidad.

2º. Corresponde al estado de avance de conocimientos; el conocimiento de variables altamente correlacionadas al fenómeno estudiado facilita la planificación de la investigación. Además, la verificación por el enfoque descriptivo de los resultados de laboratorio, permite su generalización en condiciones naturales.

3º. La elección depende de la factibilidad, ya que en ciertas condiciones, la aplicación de uno de los métodos puede ser imposible. En este caso, se recurre a verificaciones indirectas.

Entre ambos enfoques, existe una gama de variantes que se califican de enfoques "*quasi experimentales*". Este enfoque se usa cuando hay una gran dificultad de controlar todas las variables del ambiente o la imposibilidad de seleccionar aleatoriamente las unidades de muestreo. Se usa también para verificar el efecto de un factor en condiciones enteramente naturales.

En el ejemplo bajo análisis aquí, los dos enfoques pueden ser usados (descriptivo y experimental). El estado de la situación de un conjunto de lagos más o menos afectados por la lluvia ácida puede ser proporcionado por el enfoque descriptivo, dando una imagen fiel de las características bio-fisico-químicas de los lagos y hacer resaltar las relaciones entre diversas características (estado de las poblaciones y pH). El enfoque experimental puede ayudar a aislar el efecto del pH sobre otros factores.

El enfoque descriptivo puede ayudar también a determinar el efecto combinado con otros factores y valorar si el efecto se invierte, se nulifica, se adiciona o se multiplica.

El enfoque *quasi* experimental, podría consistir en las características de un lago por manejos particulares o por la agregación de cal (para aumentar el pH), sin poder controlar los otros factores.

Por otra parte el estudio de la estadística necesita la comprensión de siete conceptos fundamentales: 1) el elemento, 2) la población estadística, 3) la muestra, 4) el muestro al azar, 5) la variable, 6) las cifras significativas y 7) análisis estadísticos (Scherrer, 1984).

1) El elemento o unidad de muestreo es una entidad concreta como un individuo, un sujeto, un objeto o abstracta como una asociación vegetal, un punto en el espacio y el tiempo, una relación conductual, entr otros.

2) La población estadística es el conjunto de elementos. Se puede definir como la totalidad de observaciones individuales sobre la cual se hacen inferencias estadísticas en un área muestreada con límites espaciales y temporales claramente identificados, más específicamente **"es la colección de elementos que poseen al menos una característica común y exclusiva que la permite identificar y distinguir de cualquier otra y de la cual se puede extraer una muestra y sobre la cual se pueden hacer inferencias, deducciones y conclusiones estadísticas"**. La población puede ser finita o infinita, dependiendo del número de elementos que las componen; por ejemplo es infinito el número de experiencias sobre el reflejo en el comportamiento, puesto que se pueden repetir un número infinito de veces en condiciones parecidas. En cambio, el estudio de las poblaciones naturales son consideradas finitas puesto que las condiciones en las que se encuentran al momento y el lugar no son reproducibles a voluntad a menos que se instalen las condiciones en varios años antes (por ejemplo una decena de años). Bajo estas condiciones la población estadística a estudiar debe estar bien definida a partir de la muestra para hacer inferencias a la población a juicio del biólogo.

3) Muestra. **"Es un fragmento tomado de la población para juzgar o analizar a la población. La muestra debe ser tomada de manera particular dependiendo del estudio que se quiere analizar y debe ser representativa de la población"**.

4) El **muestreo al azar.** Para que los resultados puedan ser generalizados a una población estadística, la muestra debe ser representativa de esta; es decir, debe reflejar fielmente su composición y su complejidad. Sólo el muestreo al azar asegura la representatividad de la muestra por lo que su diseño (tipo de muestreo para obtenerla) debe ser cuidadosamente realizado.

5) Las **variables**. Es una característica medida observada sobre cada uno de los elementos de la muestra, o bien entidades definidas asociadas a las unidades de muestreo, ejemplo: la temperatura interna en el cuerpo de una rata, su talla, peso, etc. Se habla entonces de **variables propias**. Cuando se habla de un componente particular de su ambiente como el alimento disponible, la temperatura ambiente etc, se habla entonces de **variables asociadas** puesto que no son medidas, sobre el elemento propiamente dicho. También entre las características se encuentran las **variables cualitativas o discretas**, que son aquellas que no pueden ser medidas (sexo, raza, especie, estado civil), también se les llama atributos y las **variables cuantitativas o continuas**, que son aquellas que pueden ser medidas como la altura, el ancho, el peso, etc, de un elemento u objeto de estudio.

6) Las **cifras significativas**. Cada variable cuantitativa presenta un dispositivo de medida y su disposición fija las cifras significativas de un dato. Estas cifras son las cifras exactas que constituyen un número sin incluir los ceros identificada por tener un punto en donde se conforma la unidad de medida.

Las escalas numéricas en las que son representadas las variables son la cardinal, ordinal y nominal.

Las ***escalas nominales*** se utilizan para categorías de variables que se relacionan utilizando nombres o números nominales. La escala nominal establece una relación de equivalencia y todos los eventos u objetos que pertenecen a una categoría tienen una característica igual. El número de eventos que pertenecen a cada categoría se denomina **frecuencia**. Un ejemplo típico de una variable expresada en una escala nominal es el sexo en una colonia de animales de la misma especie, en donde sólo existen dos posibles valores.

La ***escala ordinal*** se utiliza cuando las categorías pueden ser ordenadas con base en algún criterio particular cuya propiedad de orden nos permite establecer relaciones tales como: mayor que, igual a, menor que, etc. La escala ordinal exige un ordenamiento antes de dar inicio con algún tipo de medición.

La ***escala cardinal*** es más refinada que la nominal y ordinal; y su mayor grado de exactitud la hace la expresión de uso generalizado en las ciencias exactas. Al utilizar variables expresadas en la escala cardinal podemos tener la siguiente subdivisión:

Escala de intervalo la cual utiliza el cero como un valor arbitrario, siendo imposible establecer razón por cociente entre las cantidades que con ellas se miden. Ejemplo: Se colocan los especimenes A y B sobre una mesa haciéndolos coincidir en su parte anterior; se le hace una incisión en la parte media a uno de ellos y se miden ambos haciendo coincidir el cero de la regla de medición con la incisión. El resultado es A = 40 mm y B = 20 mm. Se establece la diferencia 40 – 20 y se dice que el

espécimen A es 20 mm más grande que el espécimen B. Sin embargo dichas medidas no permiten establecer la razón 40 / 20 = 2, y decir que el espécimen A mide el doble que B.

La **escala de razón o cocientes** es aquella en donde se utiliza el cero como valor real siendo posible establecer proporcionalidades entre las variables que son medidas con esta escala. En el ejemplo anterior si utilizamos la longitud total de cada espécimen en lugar de hacer coincidir el cero con la incisión, entonces con el cero real la longitud A = 90 mm y B = 70 mm establecemos la razón por cociente 90 / 70 = 9/7, y se puede decir que la longitud B es 7/9 de A.

7. **Análisis estadísticos**. Un buen estimador, es una expresión matemática que mide a partir de datos de la muestra un parámetro de la población estadística (*p.e.*, media, mediana, moda, porcentaje, etc.)

2.1.5.3 Selección del elemento

En la mayoría de los estudios, el elemento sobre el que se hacen las mediciones corresponde a un individuo o a una entidad biológica claramente definida, como: un nido, un huevo, una colonia, etc.

Bajo ciertas circunstancias, la situación es más compleja y se debe elegir entre diversas unidades que responden más o menos adecuadamente a los objetivos fijados y a las restricciones encontradas, se trata por ejemplo de unidades definidas espacio-temporalmente, que necesitan la determinación de la talla en términos de superficie y del intervalo de tiempo y de la condición natural o artificial de los elementos.

La etapa de planificación de muestreo es la parte más discutida de la investigación. No es esencial que el elemento tenga un significado intrínseco, pero es muy importante que el par “elemento-variable”, es decir la medida de una variable en un elemento lo tenga. Así, en el caso de la elección del tamaño de una parcela de un metro cuadrado para estudiar la densidad de grandes ungulados sería completamente inapropiado. De esta forma se ve que la talla del elemento es a veces importante, pero también la naturaleza o condiciones en que se encuentran las unidades es importante: Por ejemplo, no es lo mismo medir factores fisiológicos en individuos estresados por alguna causa, que en otros en condiciones normales.

Como el elemento es el objeto de selección para formar la muestra, debe ser accesible, numerable y posible de colectar en una cantidad suficientemente grande. Cuando esto sucede, los errores aleatorios de estimación y el sesgo de ciertos estimadores son tanto más pequeños que la talla de muestra es grande.

2.1.5.4 Selección de la población estadística

Este proceso no siempre es muy claro. Las decisiones sobre la población estadística, son muchas veces intuitivas. En algunas ocasiones esta elección está predeterminada por la naturaleza del mismo estudio. Así en manejo de recursos naturales, las poblaciones estadísticas, se confunden con las unidades de manejo definidas administrativamente y no pueden ser modificadas por razones de jurisdicción. Por ejemplo: recursos pesqueros altamente migratorios como el atún, la sierra, etc., que atraviesan dos o más países en su migración. En biología experimental, no es el sujeto sino el material biológico el que condiciona con frecuencia la población estadística, ya que las poblaciones biológicas de animales de laboratorio se confunden habitualmente con las poblaciones estadísticas.

Para la selección de la población estadística hay ciertas reglas que sirven de guía al investigador para determinarla.

1. Validez externa, se refiere a que las conclusiones son tanto más universales mientras más extensa es la población. Aquí, dos factores importantes limitan su talla en estudios descriptivos y son:

a) Las dificultades de muestreo en el medio natural y,

b) La pérdida de resolución o precisión debido a un esfuerzo de muestreo (talla de muestra) muy débil, en relación con la variabilidad de la población, la cual puede crecer con su extensión.

Así, el estudio de las características del hábitat de una especie, se hará tomando a la población estadística como el área de repartición de la especie. Sin embargo, esta es generalmente demasiado extensa para ser cubierta por un solo proyecto.

2. Debe tener una escala de observación elegida. Según los objetivos que se hayan fijado, la escala de observación puede variar, mientras esta sea más grande, también lo será la población estadística.

3. Debe ser empírica, es decir que entre más original y compleja es la problemática a estudiar es importante trabajar sobre una población estadística conocida y fácil de estudiar; y recíprocamente, mientras la problemática sea más común o conocida, la población estadística deberá ser particular y poco conocida, para crear información nueva que pueda abrir nuevas perspectivas de investigación.

4. Finalmente si es mayor la amplitud de variación de los factores estudiados en el seno de la población estadística, más será tangible su efecto sobre el fenómeno estudiado. Por ejemplo, no se puede evidenciar en un estudio descriptivo el efecto del pH de los lagos sobre el éxito de la reproducción de peces, si todos los lagos tienen aproximadamente el mismo pH (no hay variación). Sin embargo, encontrar esas variaciones implica extender mucho la escala tiempo y espacio, lo cual es muy costoso y se requiere mucho personal y recursos para llevar a cabo estos estudios.

2.1.5.5 Selección de variables

Para alcanzar los objetivos planteados o verificar las hipótesis de trabajo establecidas, es necesario identificar las **variables dependientes** que describen diferentes facetas del proceso en estudio, así como las **variables explicativas** que concretan mejor los factores de los cuales se presume el efecto. Para esto es importante considerar los 5 criterios siguientes:

1. **Sea completa**, una variable lo es, si permite describir todas las situaciones posibles. Esta situación es solucionada por una buena selección de variables tomadas de un gran número de otras, que bajo circunstancias diferentes lleven al mismo resultado.

2. **Pertinencia**, una variable lo es, si la información que aporta, conduce a la solución del problema. Si el estudio es demasiado general, todas las variables se hacen pertinentes y el estudio se hace irrealizable, por otro lado, si se conservan todas las variables para las cuales no hay razón fundamental para rechazarlas y no aquellas por las que se tiene una buena razón de conservarlas, se corre el riesgo de fallar el objetivo de la operación.

3. **Independencia**, consiste en la evaluación al agregar otra variable en función de la información adicional que aporta. Por ejemplo se hace inútil medir el diámetro y el perímetro de un nido, ya que la información es redundante.

4. **Validez**, consiste en determinar en que medida la variable retenida corresponde al concepto estudiado.

5. **Competencia**, toda investigación se hace con recursos limitados, por lo que con este criterio se trata de determinar si los recursos asignados a la medida de una variable, no impedirá la realización de otras partes importantes del estudio.

En el contexto de un estudio experimental sobre los efectos de una disminución del pH sobre la trucha, Menendez (1975), midió regularmente en estanques experimentales y testigos (control) las variables ambientales siguientes: el pH, la alcalinidad, la acidez, la clorinidad, la dureza del agua, el oxígeno disuelto y la temperatura. Estas variables son **pertinentes** en la medida en que su efecto sobre la dinámica de poblaciones es probable. Son **válidas**, ya que reflejan ciertas condiciones fisico-químicas del agua. Son **incompletas**, ya que no hay información sobre metales pesados cuyo efecto sobre la sobrevivencia ha sido probada. Son no redundantes (**independientes**), ya que todas proporcionan a priori una información diferente. En cuanto a la **competencia**, no existe, ya que el costo de estas mediciones es mínimo. En cuanto a variables dependientes, se midieron la viabilidad de los huevos puestos calculando la proporción de ellos que presentan un desarrollo neural en el 12º día de incubación. La validez de esta variable se explica, ya que ha sido demostrado que el retraso de esta estructura tiene un efecto sobre la sobrevivencia de los huevos o los embriones.

2.1.5.6 Selección de dispositivos de medida

Para cada una de las variables retenidas, un dispositivo de medida debe ser elaborado y debe poseer 4 propiedades principales: **fidelidad, exactitud, sensibilidad** y **eficiencia**.

Un dispositivo es **fiel** si la repetición de medidas del mismo elemento en condiciones rigurosamente semejantes, proporciona resultados idénticos. La falta de fidelidad introduce un error aleatorio que puede reducirse repitiendo la medida y conservando la media de las medidas.

La **exactitud** de un dispositivo, se refiere a la ausencia de sesgos o de errores sistemáticos.

La **sensibilidad** de un dispositivo de medida, se refiere al poder de resolución, es decir a la más pequeña diferencia de valores detectable.

El concepto de **precisión**, integra los tres anteriores. Indica el intervalo en el cual un valor exacto de una medida tiene probabilidades muy altas de encontrarse (95% generalmente). La precisión es expresada generalmente en valor relativo.

La precisión es con frecuencia expresada en valor relativo, pues el intervalo crece con la magnitud medida, pero lo puede ser también en valor absoluto.

La falta de precisión puede provenir de una falta de exactitud o de fidelidad. Sin embargo, es inútil tener un dispositivo muy sensible si es poco fiel. En ecología, es inútil contar hasta con una precisión de un individuo si es difícil de hacer el conteo. Hay que notar, que según diversos autores, la precisión no siempre integra la noción de exactitud.

La **eficacia**, corresponde a la relación de la precisión con el costo. Evidentemente, los dispositivos eficaces son preferibles, pero generalmente existe una relación directa entre la precisión y el costo, de tal manera que a eficacia igual, se dispone de diversos métodos de precisión diferente.

Se opta generalmente por un dispositivo costoso cuando la variación de los datos de un elemento a otro es muy débil en relación al error de medida. Contrariamente, se escogerá un dispositivo de medida poco preciso y poco costoso en tiempo y en dinero cuando la variación entre elementos sea muy grande en relación a los errores de medida, o mientras que los costos de la selección de un elemento sean poco costosos en relación a los de la medida.

2.1.5.7 Selección del plan de muestreo

El muestreo es uno de los aspectos tratados con más negligencia de la bioestadística. Para que los resultados sean generalizables a la población estadística, la **muestra** debe ser **representativa** de esta última, es decir debe reflejar fielmente su composición y su complejidad. Sólo el muestreo aleatorio simple asegura la representatividad.

Una muestra es aleatoria cuando cada elemento de la población tiene la misma probabilidad de ser seleccionado, esta es conocida y no nula. Entre los muestreos aleatorios más utilizados están: el aleatorio simple, el sistemático, con probabilidades desiguales, estratificado y por grados, entre otros (figura 3, tabla 1).

El **muestreo aleatorio simple** (MAS), consiste en seleccionar al azar y de manera independiente ***n*** unidades y cada una de las muestra de talla ***n***, posee la misma probabilidad de ser constituida.

Además, el MAS asegura la independencia de los errores, es decir, la ausencia de autocorrelaciones entre los datos de una misma variable, independencia indispensable para la validez de varios tests estadísticos y principalmente los análisis de varianza.

El **muestreo con probabilidades desiguales de selección** de las unidades, agrupa una serie de técnicas que consisten en tomar aleatoriamente ***n*** elementos entre ***N*** unidades caracterizadas por tener oportunidades desiguales de salir seleccionadas, pero conocidas, diferentes de cero y de suma igual a uno. La aplicación más común de este plan, es el sorteo con probabilidades proporcionales o aproximadamente proporcionales a la talla de las unidades (figura 3).

Así, si x_i es la talla o una estimación de la talla de la i-*em* unidad de muestreo, la probabilidad de selección es:

$$P_i = \frac{x_i}{N} \; ; \; N = \sum x_i$$

Hay varios procedimientos para poner en práctica este tipo de muestreo. Aquí veremos el ejemplo clásico utilizado por Caughley (1977). Para estimar la densidad y el número total de grandes mamíferos a partir de conteos aéreos, la región a inventariar fue dividida en unidades de talla variable cuya forma facilitaba la navegación y la ubicación de los animales y del cual la superficie no implicaba un exceso en la duración máxima de atención sostenida de un observador. Sobre un mapa con los límites de las unidades de muestreo, ***n*** puntos eran repartidos aleatoriamente sorteando al azar sus coordenadas geográficas. Cada unidad conteniendo un punto pertenecía a la muestra y las unidades conteniendo

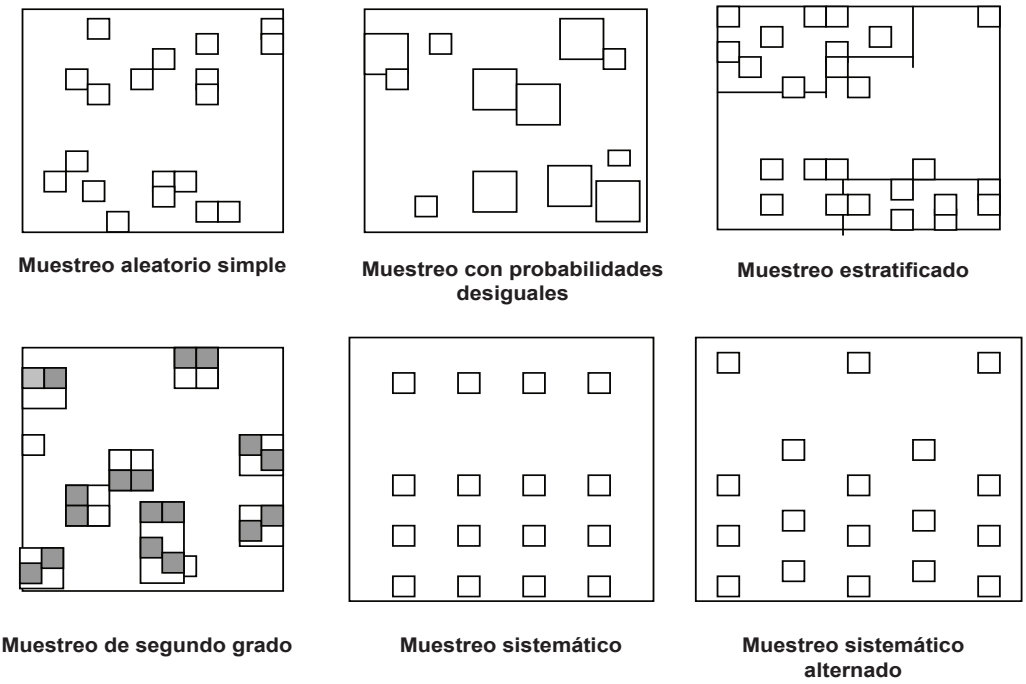


Figura 3. Distribución espacial de áreas de colecta o de medida, conforme a diversos planes de muestreo. El sistemático alternado maximiza las distancias entre estaciones y reduce las redundancias en las informaciones colectadas en el seno de poblaciones fuertemente autocorrelacionadas (Yomado de Scherrer (1984), p. 58).

varios puntos eran contabilizados varias veces. Este modo de selección constituye un muestreo con reemplazo, con probabilidades de selección proporcionales a su talla.

Este plan se aplica cuando el sorteo aleatorio simple trata con sub-unidades cuyo número varía considerablemente de un elemento a otro. Se usa también cuando la talla de los elementos varía considerablemente y que los resultados son expresados por unidad de superficie o de volumen (ejemplo cantidad de grandes mamíferos por km^2)

El **muestreo por grados** agrupa varios planes de muestreo caracterizados por un sistema ramificado y jerarquizado de unidades. Cada N unidades de la población llamadas primarias o racimos (grupos), se compone de M_i sub-unidades más pequeñas, llamadas unidades secundarias, las cuales a su vez pueden contener K_{ij} unidades terciarias y así sucesivamente.

A cada nivel, un muestreo aleatorio puede ser efectuado. Si sólo hay un MAS, se denomina muestreo de primer grado, si hay dos MAS de segundo grado, etc.

a) Muestreo de primer grado, consiste en tomar aleatoriamente n unidades primarias del total de N de la población y a medir las M_i sub-unidades, es decir todas las unidades secundarias de las n unidades primarias seleccionadas.

b) Muestreo de segundo grado, llamado submuestreo, consiste en realizar dos muestreos aleatorios. El primero de n elementos, se refiere a las unidades primarias. El segundo, contiene m_i elementos de la M_i unidades secundarias de cada unidad seleccionada, en este no todas las unidades secundarias son medidas, sino sólo las seleccionadas.

Tabla 1. Principios, ventajas e inconvenientes de diferentes estrategias de muestreo.

Tipo de muestreo	Principios	Ventajas	Inconvenientes
Muestreo aleatorio simple (MAS)	-Todos los elementos de la población tienen la misma probabilidad de ser seleccionados - Los elementos son tomados independientemente - Los elementos pueden ser o no, reemplazados en la población después de la selección. Es decir, muestreo con o sin reemplazo, el 2º es más eficaz	- No requiere un conocimiento previo de la población - La muestra es representativa de la población. - Aparte los cocientes, sus estimadores son sin sesgo - Asegura la independencia de los errores - La mayor parte de las pruebas de hipótesis y análisis estadísticos son directamente aplicables La subdivisión de la población y de la muestra en varios grupos es posible después de la colecta de datos. -Tiene una gran flexibilidad de análisis y de tratamiento	- El método de selección de elementos es poco cómodo y requiere para los elementos de las poblaciones finitas, su enumeración - No utiliza la información previa, es menos eficaz que el muestreo estratificado o por regresión (Scherrer, 1982)
Muestreo con probabilidades de selección desiguales	- Los elementos de la población tienen una talla o una importancia desigual	- El plan permite sobre-representar los elementos más importantes o los de mayor talla, sin afectar la representatividad de la muestra	- Si la literatura sobre la estimación de parámetros calculados a partir de datos surgidos de muestreo con probabilidad desigual es muy abundante, la referente a los análisis y pruebas estadísticas es muy rara
Muestreo con probabilidades proporcionales a la talla de los elementos	-Cada elemento de la población tiene una probabilidad de selección proporcional a la talla o a la importancia de las unidades, o a una estimación de esa talla. Puede ser fijada según otros criterios. - Los elementos son seleccionados aleatoriamente	- Si el denominador de una variable cociente corresponde a la talla del elemento, el muestreo con probabilidades proporcionales proporciona estimaciones no sesgadas de la media - Si la variable estudiada está correlacionada con la talla o con una estimación de la talla de las unidades, este muestreo es más eficaz que el MAS	
Muestreo por grados	-La población se compone de N unidades primarias (grupos o super-grupos), cada una de las cuales contiene M_i unidades secundarias (elementos o grupos) que a su vez pueden contener k_{ij} unidades terciarias (sub-elementos o elementos) y así sucesivamente. -a cada nivel de unidad un MAS es efectuado	-Si los grupos o super-grupos son definidos geográficamente los costos de transporte y las molestias son reducidos considerablemente. -No requiere la enumeración de todos los elementos o sub-elementos de la población. Sólo los grupos deben ser numerados -Las características de los grupos pueden ser estimados de la misma forma que los elementos de la población -La red de grupos seleccionados aleatoriamente puede ser reutilizada después, aún si los elementos que componen los grupos ya no son los mismos -El plan es una respuesta afortunada al problema de desproporciones de escala entre la talla de la unidad de muestreo (muy grande) y aquella para la que el dispositivo de medida es apropiado	-La colecta de los elementos no es independiente y los errores no podrán serlo, sólo si los elementos son distribuidos aleatoriamente en la población, lo cual es muy raro. En consecuencia, las condiciones de aplicación de varias pruebas y el de análisis de varianza en particular, corren el riego de no ser respetadas El análisis y las pruebas estadísticas aplicables a este plan quedan confinados a una literatura poco accesible, muy especializada. -Los estimadores son numerosos y variados

Tabla 1 (Continuación). Principios, ventajas e inconvenientes de diferentes estrategias de muestreo.

Tipo de muestreo	Principios	Ventajas	Inconvenientes
Muestreo estratificado	-La población generalmente heterogénea es subdividida en sub-poblaciones más homogéneas -En cada sub-población o estrato un MAS es ejecutado	-Las características de cada estrato pueden ser estimadas y comparadas entre ellos -La muestra permanece representativa de la población aún si ciertos estratos se revelan sobre-representados -El plan permite beneficiarse de ciertas situaciones favorables identificando cada situación con un estrato y modelando el esfuerzo de muestreo en función de las facilidades de cada estrato -La estratificación conlleva ganancias en la precisión -Los errores de clasificación de los elementos en los estratos disminuyen la ganancia en precisión, pero no generan sesgos si el peso de los estratos es correcto (válido para estimaciones de la población)	-Requiere el conocimiento previo del tamaño de cada estrato -Complica el tratamiento de datos reagrupados a posteriori, según un criterio diferente de la estratificación -Varios análisis y pruebas de hipótesis ya no son aplicables directamente
Muestreo sistemático	-Después de haber seleccionado al azar el 1er elemento, las subsecuentes selecciones sistemáticas de todos los elementos entre la 1a y la *p-ésima* selección se realizan a intervalos de *p*	-El plan es fácil de preparar y de ejecutar -El plan facilita el análisis y la representación de la distribución y de la variación espacial o temporal de una característica estudiada -Se presta bien para el estudio de autocorrelaciones	-Las estimaciones pueden ser muy sesgadas si hay una periodicidad en la secuencia de los elementos y si el intervalo *p* se aproxima a un múltiplo de la longitud de onda del fenómeno cíclico. -La selección no siendo independiente, el error puede no serlo tampoco. Las condiciones de aplicación de varios tests no son entonces respetadas
Muestreo no aleatorio o a criterio	-Selección de unidades representativas de la población son juzgadas por el investigador -Selección de un subgrupo de la población, la cual a partir de información es juzgada representativa, El conjunto de elementos del sub-grupo o una muestra de este último es seleccionada -Selección a ciegas	-Reduce los costos de preparación del muestreo sobre el terreno -No necesita ninguna información *a priori* -La población estadística puede ser definida a posteriori -Reduce los costos de preparación del muestreo -Cuando los elementos de la población son raros, diseminados y muy seguido, poco accesibles, este plan constituye el único enfoque técnica y financieramente realizable -Reduce el costo de preparación y de ejecución del muestreo -Es cómodo y sin riesgo importante en poblaciones muy homogéneas	-La variabilidad y el sesgo no pueden ser medidos ni controlados -La selección de una unidad depende de la percepción del observador y de la imagen mental que tiene de la representatividad de la población. En el análisis de resultados no se puede distinguir si lo que se refleja es la población o la imagen que percibió el observador -Se requiere un conocimiento perfecto del sub-grupo y de la población sino, las hipótesis a formular son muy estrictas y difíciles de cumplir -La validez de los resultados depende del grado de homogeneidad de la población y del grado de independencia de la selección

Tabla 1 (Continuación). Principios, ventajas e inconvenientes de diferentes estrategias de muestreo.

Tipo de muestreo	Principios	Ventajas	Inconvenientes
Muestreo no aleatorio con selección razonada	-Selección de elementos en función de sus características particulares a fin de maximizar la variación sobre uno o varios factores seleccionados previamente y de minimizar la variación sobre los otros -Selección de los elementos en función de su posición estratégica en la población o de ciertas de sus características, las cuales según la experiencia del investigador tienen un valor indicativo alto en la comprensión del estado y del funcionamiento de un sistema del cual se conocen los mecanismos mayores	-Permite verificar de manera cómoda y rápida la existencia o no de una tendencia lineal entre las variables explicadas y los factores escogidos con anterioridad -En conjunción con los análisis multidimensionales, este método proporciona resultados muy grandes, en función de la inversión -Para los fenómenos de los cuales se conocen los principales comportamientos, permite de dar un diagnóstico rápido y de hacerse una idea de comportamientos particulares de un sistema complejo	-El análisis termina por una respuesta positiva o negativa a la pregunta sobre la existencia de una tendencia. Toda interpretación más detallada se estrella a los siguientes problemas: La muestra no es para nada representativa de la población y la estimación de los parámetros es cargada de un sesgo importante. Aún con análisis multidimensionales es difícil detectar la estructura de la muestra, por consiguiente de disociar los resultados propios de la muestra y de la población -La interpretación de resultados es válida en la medida en que el poder indicador de los elementos seleccionados es real -Como el plan precedente, la muestra no es representativa y se hace difícil disociar el efecto de la estructuración de las propiedades de la población. Muchas interpretaciones deben ser tomadas con precaución.

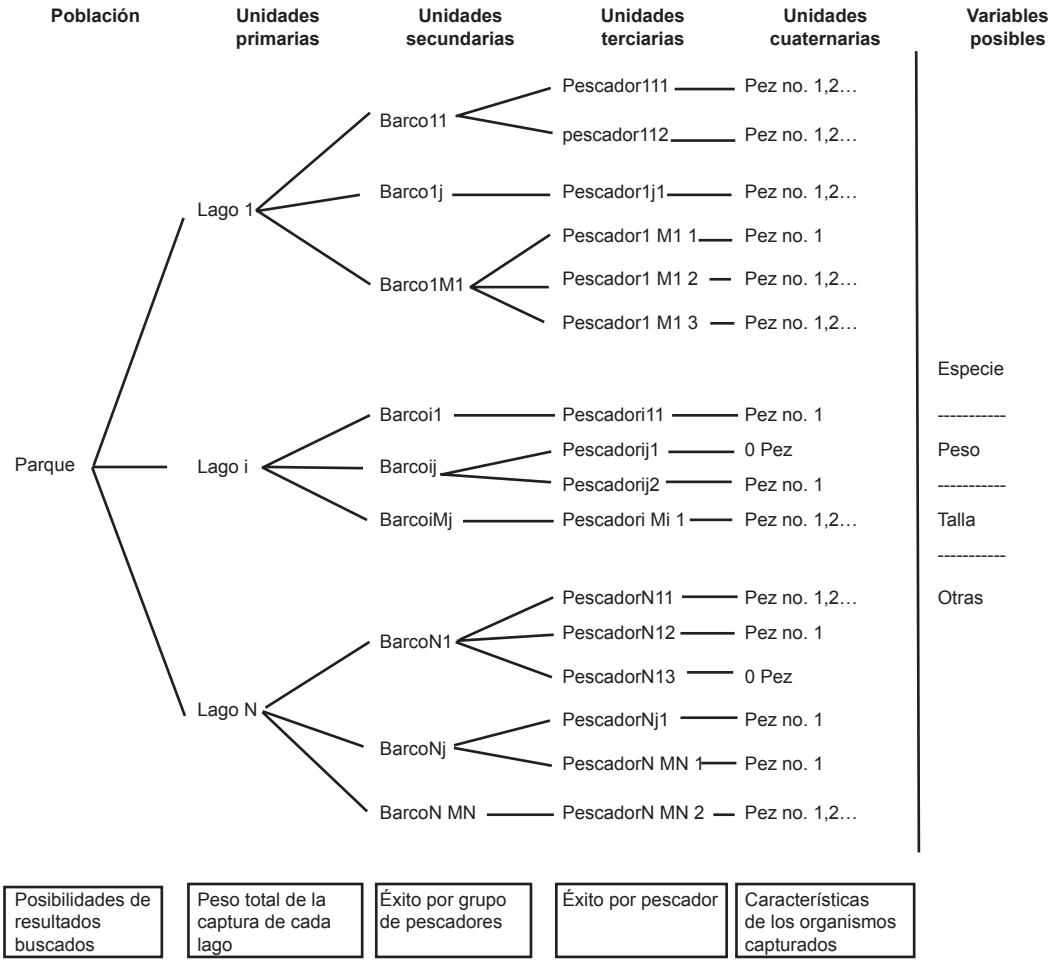
c) Muestreo de tercer grado y aún de cuarto grado, son una extensión del principio de sub-muestreo. Así, en el muestreo de tercer grado, n unidades de N unidades primarias son seleccionadas de la población, m_i unidades son tomadas de M_i unidades secundarias que componen cada unidad primaria seleccionada, finalmente, k_{ij} unidades terciarias son retenidas aleatoriamente de las K_{ij} que componen cada unidad secundaria seleccionada. Las unidades primaria, secundaria y terciaria, no son obligatoriamente de la misma talla.

El muestreo por grados se impone cuando se está en la imposibilidad de inventariar a todos los elementos de la población, para seleccionarlos aleatoriamente; pero es posible enumerarlos dentro de grandes bloques seleccionados al azar. También se aplica cuando se está en imposibilidad de medir los elementos de toda la unidad. Así, en la medida de la disponibilidad de alimento, en las áreas de pastoreo de grandes mamíferos, sólo puede realizarse por sub-muestreo. Además los estudios ecológicos que tratan sobre especies de talla muy diferente necesitan con frecuencia un sub-muestreo, ya que los animales de talla pequeña son generalmente muy numerosos para ser enumerados.

El muestreo por grados se impone también cuando los resultados son buscados en los diferentes niveles de unidades. Así para las estadísticas sobre la pesca deportiva, los resultados por lago (unidades primarias), barco (unidad secundaria), por pescador (unidad terciaria) y por pescado capturado (unidad cuaternaria), interesan al administrador del recurso (tabla 2).

El muestreo estratificado, consiste en subdividir una población heterogénea en sub-poblaciones o estratos más homogéneos interiormente, exclusivos y exhaustivos.

Tabla 2 . Ejemplo de sistema ramificado de unidades (tomado de Scherrer (1984), pág 55).



La población heterogénea de efectivos ***N*** es dividida en ***k*** estratos de talla N_h de tal suerte que: $N = N_1+N_2+.....+N_{h+}....+N_k$. Una muestra independiente es después seleccionada en cada uno de los estratos de manera aleatoria.

Los criterios de estratificación dependen del objetivo que se busca. Si se quiere optimizar la precisión de una estimación, el mejor criterio será entonces la variable mejor correlacionada con la variable estudiada y para la cual se puede tener acceso a los elementos de toda la población. Si el objetivo es efectuar comparaciones entre diferentes categorías y buscar estimaciones para el conjunto de la población, el criterio será el mismo con el que han sido formadas las categorías.

La estratificación se impone cuando los resultados son buscados a nivel de cada sub-población y de la población entera. Es indispensable cuando por diferentes razones el esfuerzo de muestreo no puede ser constante. Finalmente, es muy eficaz para mejorar la precisión de las estimaciones sin aumentar el esfuerzo de muestreo.

El **Muestreo sistemático**, consiste en seleccionar al azar un ***i-ésimo*** elemento situado entre el primero y el ***p-ésimo*** de la población y después a seleccionar sistemáticamente el $(i+p)^e$, $(i+2p)^e$, $(i+3p)^e$ $...(i+(n-1)p)^e$ elemento de la población. Los rangos de ***n*** unidades son así en progresión aritmética cuya

base es un número aleatorio ***i*** y la razón un número ***p*** calculado de tal manera que la muestra se reparte uniformemente sobre toda la población.

Este plan se aplica fácilmente cuando los elementos de la población son fácilmente accesibles, en número conocido, dispuestos los unos enseguida de los otros o casi. El plan es muy apropiado para estudiar las variaciones espaciales y temporales, ya que se presta bien a representaciones cartográficas o esquemáticas de los resultados. Permite estudiar fenómenos de autocorrelación ya que las unidades seleccionadas a intervalos regulares no son seleccionadas de manera independiente, sino que contrariamente, no asegura la independencia de los errores; es decir, la independencia de las variaciones no controladas de los elementos en estudio, la independencia es un requisito para ejecutar algunas pruebas estadísticas.

Si bien el plan permite el estudio de fenómenos periódicos, proporciona una imagen deformada de la realidad tan pronto como la razón ***p*** (intervalo entre dos elementos) se aproxima a un múltiplo de la longitud de onda de las variaciones del carácter estudiado (Fig. 4).

Otros planes de muestreo son tratados en algunas obras: Cochran (1977), Sukhatme y Sukhatme (1970), Som (1973), etc., se trata esencialmente de muestreo por regresión, por cociente, entre otros. Además, una multitud de planes pueden ser creados a partir de una combinación de los planes elementales. Por ejemplo, un muestreo de tercer grado con estratificación de unidades primarias seleccionadas con probabilidad proporcional a su talla y sin reemplazo, con unidades secundarias seleccionadas con probabilidades iguales y sin reemplazo y unidades terciarias seleccionadas según muestreo sistemático constituye un plan nada teórico, sino muy práctico (Scherrer, 1982).

Otros planes muy utilizados son: **Muestreo a criterio** y el muestreo a **selección razonada**. El **muestreo a criterio**, se basa fundamentalmente en el criterio o conocimiento del investigador para seleccionar los elementos de la muestra, agrupa para esto tres prácticas principales: muestreo de **elementos representativos** o típicos, muestreo de una **sub-población representativa** y **muestreo a ciegas**. Para el primer caso, se realiza una idea del elemento típico (retrato hablado) y el muestreador, selecciona elementos de la población que se asemejen al elemento típico. En ecología, muchas estaciones de muestreo son seleccionadas de esta manera, en función de la representatividad de un ambiente. Tiene

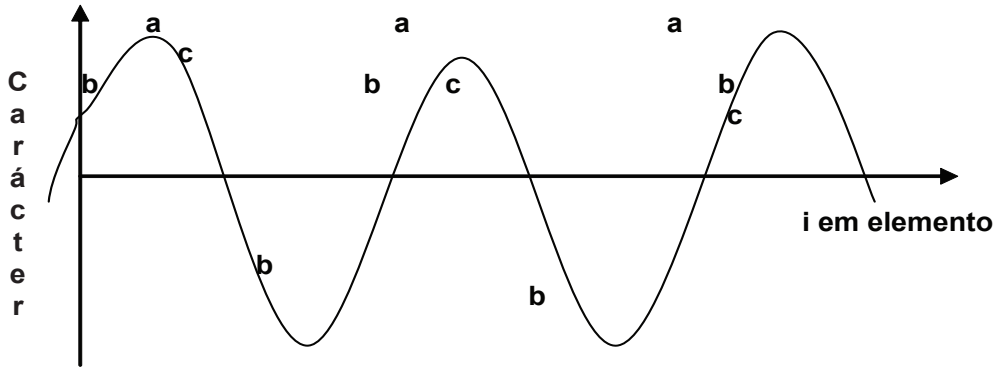


Figura 4. Muestreo sistemático efectuado en una población presentando variaciones periódicas. Puntos a: *p=k*, puntos b: *p=9k/16*, puntos *c=k/2*; *p*: razón o paso de muestreo y *k*: longitud de onda de variación del carácter estudiado (tomado de Scherrer (1984), pág 57).

el inconveniente de que la imagen puede cambiar según la época y otros factores. Para el segundo caso, la búsqueda de un grupo representativo de la población, requiere un conocimiento previo de la población y de la sub-población. Los resultados a nivel de sub-población sirven como indicador de la evolución del comportamiento de la población. Se usa para minimizar los desplazamientos en zonas poco accesibles, o demasiado vastas, o por toda razón técnica, financiera o humana. Sin embargo, fuertes hipótesis sobre la semejanza con la población son determinadas y nada permite su verificación. El tercer caso consiste en tomar unidades a ciegas, es muy cómodo, económico y relativamente confiable si la población es suficientemente homogénea. Sin embargo, si existen ciertas heterogeneidades, no es posible detectarlas de esta forma, y entonces, se puede incurrir en sesgos.

M**uestreo a selección razonada**, consiste en seleccionar diferentes unidades en función de ciertas características, permite obtener una estructura particular y buscada en el seno de la muestra. En general se buscan dos tipos de estructuras: el primero es un gradiente de valores sobre una o varias variables, con un mínimo de variación sobre las otras. Por ejemplo, el comportamiento (correlación) de la abundancia de una comunidad animal o vegetal en función de una o varias variables, estas forman parte del ambiente del medio donde habitan los organismos estudiados. Permite a bajo costo verificar el grado de significatividad de la correlación entre variables; es decir, si la tendencia al incremento o la disminución de una variable puede ser fruto del azar, o bien si la tendencia está ligada directa o indirectamente al gradiente de valores de la o las variables independientes *x*, obtenidas por la selección particular de los elementos.

Cualidades de la muestra, la **representatividad** es la primera cualidad. La segunda se refiere a la **talla**, es decir al número de ***n*** elementos que la compone, la cual debe ser tan grande como sea posible. La precisión y la robustez de la mayoría de pruebas estadísticas crecen en función de ***n***; además el sesgo introducido por la mayoría de los estimadores decrece cuando ***n*** se incrementa. Sin embargo, la **ley de rendimientos decrecientes** impone un límite al aumento de ***n***, ya que la precisión crece con la raíz cuadrada de ***n***; así la ganancia en precisión se hace rápidamente poco importante con el aumento de unos cuantos elementos suplementarios. La tercera cualidad se refiere a la **eficacia**, la cual se evalúa en términos de la precisión obtenida por unidad de tiempo o dinero invertidos en el muestreo.

A veces se habla de **poder de resolución** de un muestreo (Frontier, 1982). Se aplica generalmente al muestreo sistemático y traduce la capacidad de este para detectar las fluctuaciones próximas. Esta capacidad depende de la razón o del paso, es decir del tramo que separa dos elementos muestreados, el cual esta ligado al número de elementos ***n***. Se parece al concepto de precisión.

2.1.5.8 Selección del plan experimental

El principio de un experimento consiste en remplazar el sistema complejo de relaciones causales que se presentan en la naturaleza, por sistemas muy simplificados en los cuales un solo factor a la vez presumiblemente determinante puede variar. De manera general, el método experimental puede aplicarse tan rápido como uno pueda fijar las condiciones del sistema a modificar y en donde una sola variable entre ellas de manera progresiva, reversible y repetitiva se pueda evaluar. Evidentemente, varios campos de la biología no se prestan a la experimentación. Así los mecanismos de funcionamiento de los ecosistemas forestales y lacustres son difíciles de analizar, porque a esta escala no se pueden fijar las condiciones del sistema y hacer variar a voluntad y de manera reversible y repetitiva un factor como la temperatura o el fotoperiodo. En este caso se aplicará el enfoque **quasi-experimental,** o a la veri-

ficación indirecta. La tabla 3 muestra los principales puntos comunes y divergencias de los enfoques descriptivos, experimentales y quasi experimentales.

Algunos criterios han sido utilizados para admitir relaciones de causalidad. El método de las **diferencias** consiste en observar en varias situaciones idénticas bajo todas las relaciones a excepción de un **factor *X***, las diferencias aparecen sobre el plan del fenómeno **estudiado *Y***. Las desviaciones observadas en ***Y*** son debidas al tratamiento experimental ***X***, si y solo si ningún otro factor varía de manera incontrolada de una situación a la otra.

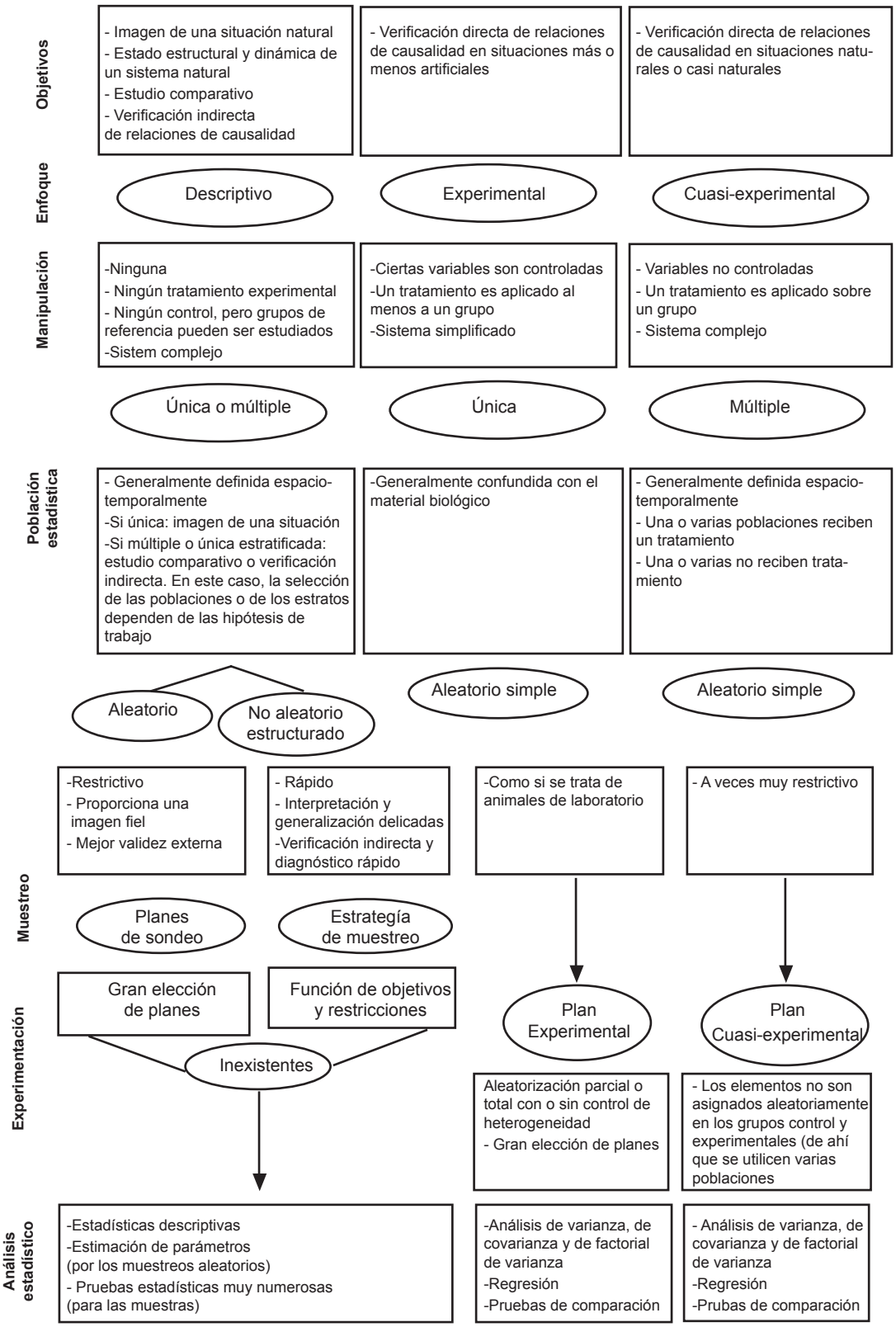
Este método es aceptado universalmente, utiliza diversos medios para asegurar la perfecta similitud de situaciones. El **primero** consiste en **eliminar** del campo experimental los factores que pueden ser suprimidos, es decir, cuyo valor puede ser reducido a cero. Así en el ejemplo sobre el efecto de una disminución del pH sobre el éxito de la reproducción de las truchas, el cloro del agua ha sido suprimido para redimensionar el efecto. El **segundo** se aplica a los factores cuyo efecto **no puede ser eliminado**, pero se puede controlar el nivel de la naturaleza de sus manifestaciones. Consiste en conservar una perfecta constancia de sus factores en cada una de las situaciones. En el ejemplo precedente, para suprimir el efecto de los desplazamientos de los investigadores o de todo estimulo visual particular, los acuarios deberán estar rodeados de una pantalla uniforme, impidiendo todo contacto visual con el exterior. La influencia de este factor es entonces mantenida constante. Sucede lo mismo con la radiación luminosa para cada acuario por tubos fluorescentes.

Para las variables propias, es decir las que caracterizan al elemento propiamente dicho, como **la edad, el sexo, la talla, el peso** de las truchas, a lo inverso de la luz, de la temperatura, que caracterizan el ambiente, toda investigación de constancia sólo se puede hacer bajo una reducción de la generalización de los resultados. Efectivamente, el control de los factores no solo limita el alcance de los resultados a situaciones particulares, sino que, además, si sólo se usan truchas de una edad o sexo determinados, los resultados se referirán exclusivamente a esa población.

Para evitar tales restricciones que pueden afectar considerablemente la validez de los resultados, se utiliza un **tercer medio** de homogeneización o de igualación de las situaciones: **la aleatorización esto significa, introducir el efecto del azar**. Consiste en distribuir **aleatoriamente** los elementos de la muestra seleccionados **aleatoriamente** de la población estadística en diversos grupos correspondientes a cada una de las situaciones. Los efectos de la multitud de factores que pueden actuar sobre el fenómeno se encuentran así repartidos de una manera sensiblemente uniforme sobre el conjunto de grupos. Aplicado al ejemplo anterior, la **aleatorización** se refiere por ejemplo a una asignación aleatoria de truchas en los diferentes acuarios correspondientes a los diferentes niveles de pH probados. La **aleatorización** se puede también aplicar a factores externos. Así, para asegurar que la composición de la alimentación provista a cada grupo sea idéntica, los diversos componentes pueden ser mezclados mecánicamente hasta que el conjunto sea homogéneo. Esto, corresponde a una **aleatorización** mecánica.

Otro medio de asegurar la equivalencia o constancia de las situaciones, o el equilibrio de los grupos es el apareamiento de grupos (matching), consiste en suprimir la independencia de las diversas sub-muestras, es decir de los diversos grupos. En el apareamiento, se modifica la probabilidad de seleccionar los elementos apareados. Finalmente, los sorteos en el seno de un mismo grupo son siempre independientes, pero de un grupo a otro las relaciones de dependencia aparecen con el apareamiento. Este medio de equilibrar, puede ir hasta la designación del mismo elemento en dos grupos diferentes; se trata por

Tabla 3. Comparación de características de los enfoques descriptivo, experimental y cuasi-experimental.



ejemplo, de los mismos elementos antes y después de un tratamiento. El apareamiento puede también tratar sobre los miembros de una misma camada, de una misma familia, de una misma colonia, de una misma asociación, etc.

Los dos primeros medios presentan la ventaja de reducir la variabilidad intra e intergrupos y de esta manera se incrementa la sensibilidad del plan experimental. El poder de detección de pequeñas diferencias provocadas por el tratamiento es elevado; pero, contrariamente las situaciones se hacen muy artificiales y entonces el alcance de los resultados se limita a estas condiciones.

El tercer medio presenta las ventajas e inconvenientes inversos que los precedentes. Más se hace uso de la aleatorización, más se coloca uno en condiciones naturales; pero desafortunadamente la variabilidad intra e intergrupo tenderá a enmascarar el efecto del tratamiento. Para reducir el error o la incertidumbre que resulta en los resultados de la variable o fenómeno a estudiar, hay que aumentar la talla de la muestra (grupos). Sin embargo, los altos costos que esto puede implicar resulta a veces imposible incrementar la talla de muestra.

El cuarto medio reduce la variación intergrupo natural y así, incrementa la sensibilidad del plan experimental; desafortunadamente su aplicación no puede ser generalizada.

La concepción del plan experimental, como el de toda investigación necesita la satisfacción de tres criterios: de validez general, de validez interna y de validez externa.

a) Criterio de validez general, corresponde a la capacidad del plan de r**esponder específicamente a la pregunta hecha**, o de probar únicamente la hipótesis a verificar. No es raro que un plan satisfaga los criterios de validez interna y externa, provea una respuesta clara, pero no responder exactamente al problema o pregunta planteada. Para esto hay que proponer o modelar el problema, de tal manera que la prueba estadística elegida responda a la pregunta planteada.

b) Criterio de validez interna, se refiere a la capacidad del plan de **prevenir la emergencia de hipótesis rivales**, que proporcionarían varias explicaciones posibles a los resultados encontrados. Este criterio puede ser satisfecho por dos mecanismos: las técnicas de equilibrio de grupos y el empleo de grupos testigos. Estos, en oposición a los grupos experimentales se refieren a situaciones en las que la variable independiente (factor estudiado) tiene con frecuencia un valor nulo (cero). Los grupos testigos juegan dos papeles: el de comparar, a fin de establecer las diferencias entre los elementos tratados y no tratados, y el de **evidenciar** hipótesis rivales aislándolas y comparando sus efectos.

c) El criterio de validez externa, se refiere a las posibilidades de **generalización**, **las cuales dependen de tres factores: la talla de la población estadística, el grado de utilización de medios de equilibrio 1 y 2 (aleatorización y apareamiento), y la representatividad de la muestra.**

El **primer plan** experimental es el más simple consiste en constituir dos grupos, uno experimental y otro testigo o control, por aleatorización o eventualmente por apareamiento. El tratamiento es aplicado al grupo experimental, las respuestas son medidas y después comparadas (Tabla 4). En agronomía, se trataría por ejemplo de parcelas seleccionadas al azar y divididas en dos grupos. Sobre el primer grupo, se aplica un fertilizante determinado, pero no sobre el segundo: la productividad en los dos grupos es medida y comparada.

El **segundo plan** consiste en formar cuatro grupos: tres testigos y uno experimental. Los dos primeros grupos habitualmente son formados por aleatorización y los otros dos son apareados a los precedentes por la conservación de los mismos elementos. Dos grupos testigos son medidos antes de la aplicación del tratamiento, y los dos restantes lo son después del tratamiento, después las desviaciones observadas sobre los elementos con tratamiento son comparadas con los de los elementos sin tratamiento (Tabla 5).

Este plan puede ser utilizado para probar la eficacia de un compuesto vitaminado suministrado a gallinas para mejorar su producción de huevos. Las gallinas serían seleccionadas al azar de una población dada, luego repartidas aleatoriamente en uno de los dos grupos iniciales. Su producción normal sería observada durante un cierto periodo y los compuestos vitaminados serían agregados al alimento de uno de los grupos. Después de un cierto tiempo, las medidas de productividad se harían para los dos grupos.

La eficacia de un tratamiento depende casi siempre sobre su valor cuantitativo o cualitativo. En la mayoría de experimentos, el factor estudiado es probado con diferentes niveles para encontrar eventualmente los umbrales inferior y superior, así como el nivel óptimo. Un plan así es esquematizado en la tabla 6, se puede denominar **plan con un factor, (g+1) niveles, con repeticiones, aleatorización total y sin control de heterogeneidad**. En efecto, hay ***(g+1)*** niveles a prueba en el factor estudiado; los grupos tienen más de un elemento y han sido constituidos por aleatorización. En cuanto a la ausencia de control de heterogeneidad, esta será tratada en los planes subsecuentes.

Este plan ha sido aplicado al experimento sobre los efectos de una disminución del pH sobre la reproducción de truchas; sobre el éxito de la eclosión de huevos y del cultivo de juveniles. En efecto, las truchas del estanque testigo fueron expuestas a un pH de 7.0, las del grupo experimental 1 a un pH de 6.5, las del grupo 2 a un pH de 6.0, el número 3 a un pH de 5.5 y así sucesivamente hasta el grupo número 6 con un pH de 4.0. Los valores de pH corresponden a los valores de ***X*** y las tasas de producción de huevos a los de ***Y***. Cabe notar que el valor de ***X*** de pH 7, no es nulo contrariamente a la definición de grupo control. Hay que notar que $\boldsymbol{Y_0}$, $\boldsymbol{Y_1}$, $\boldsymbol{Y_2}$, etc., representan conjuntos de valores, por lo que pueden ser representados como un vector. De esta forma, el vector ***Yi*** corresponde a las medidas efectuadas sobre los $\boldsymbol{n_i}$ elementos del ***i-em*** grupo, entonces se tiene:

$$Yi = \begin{bmatrix} Y_{i1} \\ Y_{i2} \\ \dots \\ Y_{ini} \end{bmatrix}$$

Medida del primer elemento del grupo i
Medida del segundo elemento del grupo i

Medida del tercer elemento del grupo i

Es el mismo caso con la variable ***X***.

El plan factorial, se usa para poner en evidencia la acción de varios factores, ya sea aisladamente, así como las interacciones entre ellos. Un plan de dos factores puede ser representado por una tabla de doble entrada, donde las diferentes columnas representan los diversos niveles del primer factor y las diferentes líneas corresponden a los diversos niveles del segundo factor (tabla 7). El número de celdas experimentales o testigos se eleva a (g+1) (h+1) y, si cada celda contiene un número equivalente de ele-

Tabla 4. Plan experimental (R aleatorización, A apareamiento)

Grupo	Modo de formación	Tratamiento	Medida	Análisis
Experimental		x	Y_e	
	R o A			Comparación
Testigo			Y_t	

Tabla 5. Plan experimental con medidas pre y post-tratamiento (R aleatorización, A apareamiento).

Grupo	Modo de formación	Medida antes tratamiento	Tratamiento	Medida después tratamiento	Análisis
Experimental		Y_{1t}	x	Y_{2t}	$(Y_{1t} - Y_{2t})$
	R (eventualmente A)				Comparación
Testigo		Y_{1nt}		Y_{2nt}	$(Y_{1nt} - Y_{2nt})$

Tabla 6. Plan con un factor, (g+1) niveles, con repeticiones, aleatorización total y sin control de heterogeneidad.

Grupo	Modo de formación de grupos	Tratamientos	Medida	Análisis
Testigo	Aleatorización	0	y_0	
Experimental I	"	x_1	y_1	
Experimental II	"	x_2	y_2	Comparación o regresión
...	"	...	...	
Experimental g	"	x_g	y_g	

Tabla 7. Plan factorial con dos factores y (g+1) (h+1) niveles.

		Factor A				
Factor B	Nivel	0	A1	A2	...	Ag
	0	Testigo	A1	A2	...	Ag
	B1	B1	A1 B1	A2 B1		Ag B1
	B2	B2	A1 B2	A2 B2	...	Ag B2
	...	...	...	...	...	...
	Bh	Bh	A1 Bh	A2 Bh		Ag Bh

mentos (n_{ij}), el efectivo total de la muestra se eleva a $n=(g+1)(h+1)n_{ij}$. Cada celda sólo puede contener un individuo, en este caso, la interacción de factores no puede ser calculada.

De esta forma, en el ejemplo de la disminución del pH sobre el valor de ciertos parámetros de la dinámica de poblaciones de truchas, el plan factorial se aplicaría al estudio de juveniles. En efecto, el crecimiento y la sobrevivencia de los juveniles pueden ser afectados por el pH del ambiente y por el grado de exposición de los padres. Por lo que hay que estudiar los dos factores y el plan factorial se presta para ello.

En la celda I los adultos no estaban expuestos a pH bajo, pero los huevos y los embriones están colocados en un estanque con pH=4.5; en la celda III los padres no estaban expuestos a pH=5.5 pero sus embriones si; en la celda XI, los padres estaban expuestos a pH bajo, pero no sus alevines y así sucesivamente (tabla 8). Para asegurar la aleatorización, los adultos fueron colocados al azar en acuarios con pH bajo o con pH normal y sus huevos fueron también repartidos al azar en las 6 posibilidades de concentración de pH.

El plan factorial se aplica también a varios factores, pero en ese caso, el número de celdas y obligatoriamente el número de elementos a tratar crecen rápidamente. Un plan con 4 factores y 2x3x2x2 niveles, necesitaría 48 celdas (tabla 9). Si el plan es con réplicas (repeticiones), es decir con varios elementos por celda, el número de elementos toma rápidamente dimensiones poco realizables.

Hasta aquí, no se ha tratado con ningún control de hetereogeneidad y se espera que la asignación de elementos al azar equilibrará las celdas. Desafortunadamente, hay factores de heterogeneidad que predominan en la variación de datos, como la existencia de gradientes; serán entonces necesarios más elementos para equilibrar las celdas. También para reducir la variación, sin por consiguiente reducir la talla de la población estadística, seleccionando a los elementos que poseen características restrictivas, se escogerá un plan experimental con control de heterogeneidad; esto es a través de la serie de **bloques completos**. Se trata de un dispositivo plurifactorial de los que un factor juega un papel particular. El **efecto de bloque**; un bloque está constituido por elementos de la misma naturaleza, es decir, teniendo un denominador común que aumenta sus semejanzas e incrementa la homogeneidad del bloque. De un bloque al otro las diferencias pueden ser bastante grandes.

En agronomía, un bloque puede representar una banda de terreno homogéneo en la cual la fertilidad del suelo, el drenaje y otros factores son tan homogéneos como sea posible. En ecología, se puede tratar de una parcela o de un cuadrante en los cuales las características del grupo son muy similares. La eficacia del dispositivo en bloques reposa sobre el hecho de que existen menos diferencias entre los sitios de un mismo bloque que entre los sitios de diferentes bloques. Si no es el caso, la asignación al azar sería preferible.

2.1.6 Plan Cuasi-Experimental

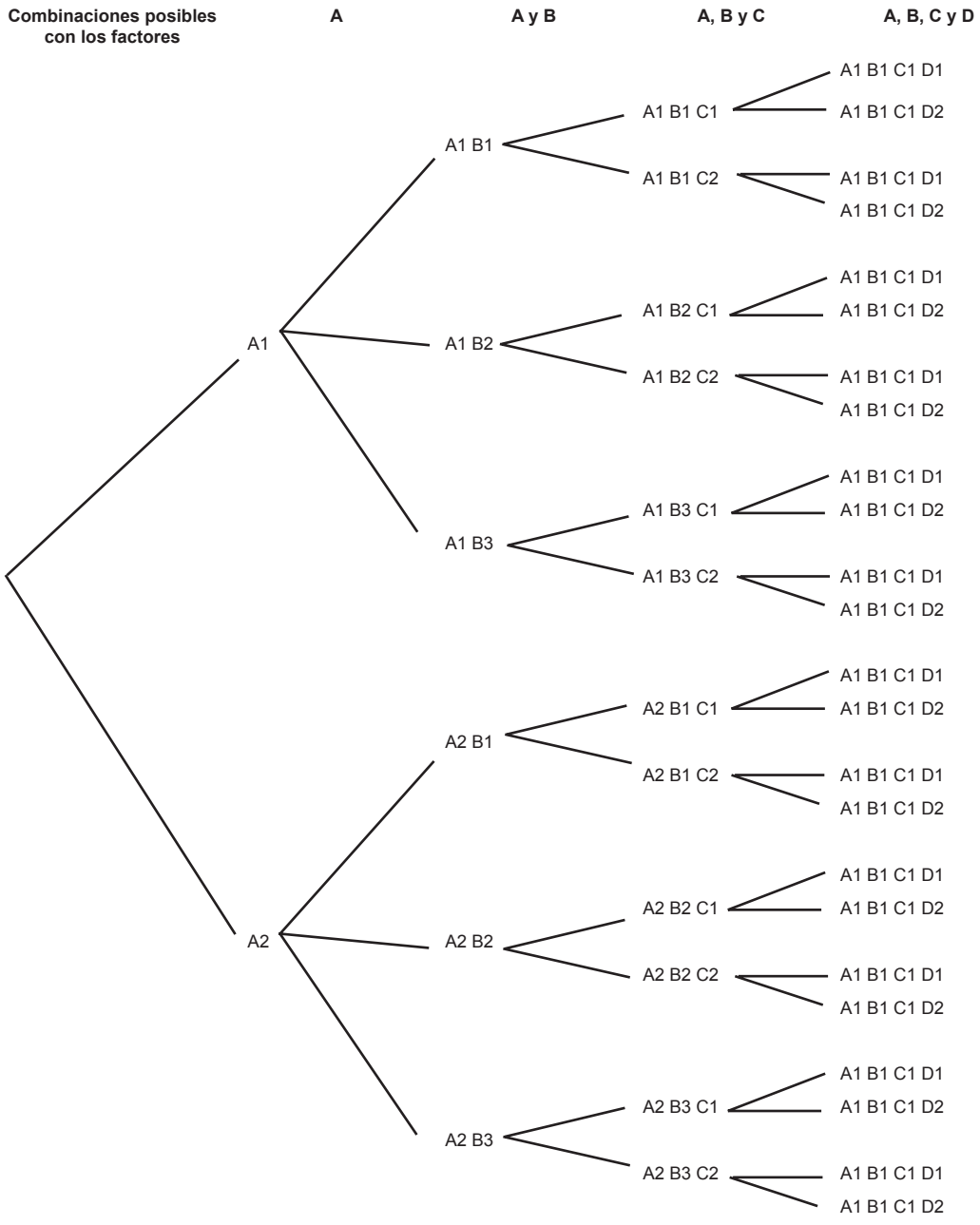
Consiste en aplicar un procedimiento experimental para analizar e interpretar datos que no llenan todas las exigencias de este. Aún si las respuestas revisten cierta ambigüedad, este enfoque a la vez experimental y descriptivo, es el único para probar directamente relaciones causa-efecto en situaciones totalmente naturales.

Entre estos modelos, uno de los más utilizados es aquel en el que los elementos no son repartidos aleatoriamente en grupos experimentales y testigos. Así, para estudiar el impacto de los desechos de

Tabla 8. Plan factorial con dos factores y 6x2 niveles, en aleatorización total.

	pH					
Padres/huevos y alevines	4.5	5.0	5.5	6.0	6.5	7.0
Padres no expuestos	I	II	III	IV	V	Testigo
Padres expuestos	VI	VII	VIII	IX	X	XI

Tabla 9. Plan factorial con 4 factores, 2x3x2x2 niveles, aleatorización total. El sistema ramificado permite de prever el conjunto de combinaciones posibles de los diferentes niveles de cada factor. El nivel 1, siempre el control, corresponde al valor cero del factor



desagües sobre la fauna béntica de un río, dos porciones correspondientes a dos poblaciones estadísticas son elegidas de un lado y otro del desagüe. Las colectas serán repartidas aleatoriamente en cada zona y son efectuadas antes y después de la operación del desagüe (figura 5). Como los elementos tratados, es decir los sitios precisos de colecta después del desagüe, pertenecen a una población estadística diferente de la de los elementos no tratados situados antes del desagüe, siempre es posible que la diferencia de las dos zonas detectadas o no antes de la puesta en operación del desagüe, explica las diferencias observadas después de la puesta en operación.

Para reducir el riesgo de que los resultados sean explicados por una hipótesis rival no deseada, un segundo modelo cuasi-experimental consiste en colectar (muestrear) una serie temporal de datos en el grupo control y otra serie del grupo experimental y en verificar la semejanza de las dos evoluciones.

Así para estudiar el efecto de las aspersiones aéreas de fenitrothion sobre la avifauna forestal de la isla de Anticosti (Canadá), Scherrer y Quellet (1973) efectuaron el estudio con intervalo de tiempo regular, con conteos de aves en los sectores tratados con el insecticida y en los sectores no tratados (tabla 10). La aspersión ha sido realizada en medio de la serie temporal de los conteos, de suerte que los efectos inmediatos y a corto plazo puedan ser detectados. La ventaja de este modelo reside en que puede detectar mejor los pasajes de migradores tardíos, los cuales pueden modificar durante un período corto el número de pájaros y así, modificar los resultados y la interpretación del experimento.

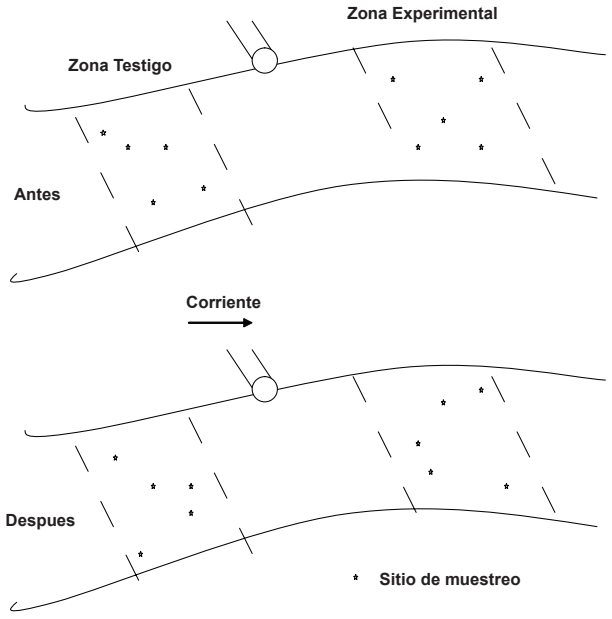


Figura 5. Plan cuasi-experimental para verificar el impacto de una fuente de contaminación sobre la fauna béntica de un río. (Tomado de Scherrer (1984), p. 83).

Tabla 10. Modelo cuasi-experimental con doble serie temporal.

	Medidas pre-tratamiento	Tratamiento	Medidas post-tratamiento
Grupo experimental	y_1 y_2 y_3 y_4	X	Y_5 y_6 y_7 y_8
Grupo testigo	Y_{t1} y_{t2} y_{t3} y_{t4}		Y_{t5} y_{t6} y_{t7} y_{t8}

2.2. Elección de Estimadores y Análisis Estadísticos

Las diversas etapas del proceso metodológico muestran que la planificación de una investigación necesita una serie de decisiones cuyas implicaciones repercuten sobre el resto del proceso, esto incluye los estimadores y los análisis estadísticos seleccionados.

Un **estimador** es una expresión matemática que mide un parámetro de la población estadística, a partir de los datos de la muestra. Para el muestreo aleatorio simple, la expresión $\bar{a} = \Sigma a_i/n$ es un estimador de la media $\bar{A}$ o μ . La elección de un estimador depende de los objetivos a alcanzar, es decir del parámetro de la población que se busca estimar (media, mediana, total, porcentaje, etc.), de la escala de variación y de la complejidad de las variables a estudiar (variables cualitativas, semi-cuantitativas, cuantitativas, simples o derivadas), del plan de muestreo utilizado muestreo aleatorio simple (MAS), estratificado, por grados, etc.). Tres propiedades los caracterizan: el sesgo, su convergencia y su eficacia.

Un estimador ***t*** de un parámetro θ es calificado como **no sesgado,** si la esperanza matemática o el valor esperado de ***t*** es igual a θ. Para una población finita, la esperanza de ***t*** es obtenida tomando la media de los valores de ***t***, calculada sobre todas las muestras posibles de talla ***n*** que pueden ser extraídas de la población. Hay que señalar que si las muestras son obtenidas por un MAS, el estimador $\bar{a}$ es no sesgado.

Un estimador es calificado como **convergente** si, por un aumento de la talla de muestra, los valores de ***t*** convergen hacia los de θ.

Un estimador es calificado de **eficiente** si, con esfuerzo de muestreo igual, proporciona estimaciones más precisas que otros.

En resumen, la obtención de una imagen fiel de alguna realidad necesita por consecuencia de muchas precauciones. La más importante consistirá en detectar entre las características anteriores, la que limita la reducción de los errores de la estimación. Conociendo el punto débil del plan de muestreo uno lo corregirá si es necesario por calibración, por aumento de la talla de muestra, modificación del dispositivo de medida o de muestreo, etc. (tabla 11).

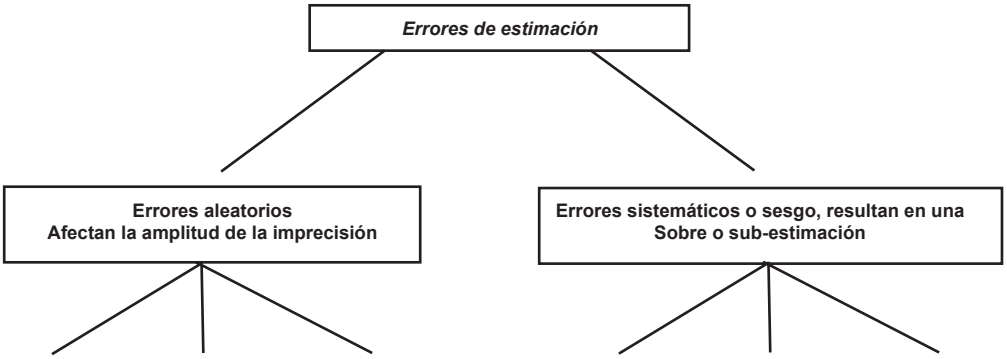
La elección de un análisis estadístico está ligado al plan de muestreo (ya sea, quasi experimental o descriptivo) o experimental. Muchos análisis de varianza y factoriales de varianza requieren una colecta de datos muy particular. Sin embargo, si varios análisis estadísticos alternativos responden a los objetivos del estudio, la elección se hará en función de 4 criterios: la pertinencia, el poder, la robustez y la compatibilidad.

El **poder** de un análisis corresponde a su capacidad de detectar pequeñas diferencias o pequeños efectos, sin por tanto aumentar la posibilidad de declarar diferentes poblaciones iguales o de calificar de activo un factor que no lo es.

La **robustez** se refiere al grado de sensibilidad del análisis al no respeto de las condiciones de aplicación. En consecuencia, la validez de los resultados surgidos de un método robusto será poco afectado por la violación de las condiciones de aplicación.

La **compatibilidad** corresponde a la posibilidad de efectuar diferentes análisis para responder a partir del mismo lote de datos a diferentes objetivos.

Tabla 11. Características de los dispositivos de medida, de los muestreos, de los estimadores y de los planes de muestreo que afectan la amplitud y la naturaleza de los errores de la estimación.



Dispositivo de medida	Muestra	Estimadores	Dispositivo de medida	Muestra	Estimadores
-Fidelidad	-Talla n	-Eficacia	-Exactitud	- Representatividad (muestreo aleatorio)	-Sesgo (ausencia)
-Sensibilidad	-Fracción de muestreo fm=n/N	-Convergencia			-Convergencia
	-Plan de muestreo (eficacia)				

2.3. Preparación del Tratamiento Informático de Datos

Una vez que los datos se hacen muy numerosos o que los cálculos se hacen complejos, es necesario hacer uso de computadoras para realizar los análisis y conocer el uso de programas estadísticos: Statistical Package for Social Science (SPSS), Biomedical Computers Programs (BMD), Statistical Análisis System (SAS), STATISTICA, XLSAT, entre otros.

2.4. Inventario de los Límites del Método

Los límites provienen de dos fuentes principales: las selecciones realizadas y las hipótesis subyacentes. Tan pronto como se procede a una selección en la elaboración de una metodología, se abren ciertas posibilidades, pero se cierran otras y como el conjunto del procedimiento necesita una serie de decisiones interdependientes uno restringe progresivamente su amplitud de interpretación segura, la cual puede llegar a ser nula por la falta de planificación.

En el marco del ejemplo sobre el efecto de una disminución del pH sobre las poblaciones de truchas, la selección de una especie particular en lugar de las poblaciones de salmónidos limita los resultados a esta especie, por lo que los resultados son de difícil extrapolación a otros salmónidos.

La otra fuente de limitación, proviene de la dificultad de obtener ciertas propiedades deseadas, como la exactitud y la fidelidad de un dispositivo de medida, la representatividad de una muestra, la validez de una variable, etc., o la imposibilidad de respetar rigurosamente las condiciones de aplicación de un método como las asociadas a la mayor parte de las pruebas estadísticas o las que aseguran el funcionamiento confiable de un instrumento. En estos casos hay que admitir ciertas hipótesis denominadas subyacentes o implícitas. Estas hipótesis son las más difíciles de descubrir ya que generalmente son

enmascaradas por la lógica del razonamiento. Sobrepasar los límites de estas hipótesis constituye la fuente de críticas negativas a las investigaciones. Si las hipótesis subyacentes son probadas como válidas, entonces la interpretación de la investigación no tiene ambigüedades. También se puede proceder a un estudio de **sensibilidad** que consiste en considerar diferentes escenarios y a establecer la validez de los resultados de cada situación. Por ejemplo, en ornitología, la densidad de una comunidad de ciertos ambientes es determinada con la ayuda de cuadrantes, los que proporcionan la mejor estimación de abundancia para las aves cantoras en periodo de anidación. Pero este método muy laborioso es con frecuencia reemplazado por el método de los índices puntuales de abundancia (IPA), el cual después de una conversión de datos, proporciona resultados de abundancia por hectárea. Desafortunadamente este método tiende a subestimar las fuertes densidades por efecto de saturación. Si se compara la abundancia de un medio M_1 obtenido por IPA con la de otro medio M_2 obtenido por cuadrantes, se requiere prever 3 escenarios posibles:

- La abundancia de M_1 es superior a la de M_2, el resultado es válido ya que el IPA sólo puede subestimar.

- La abundancia M_1 es igual a la de M_2 , para las densidades débiles el resultado es válido; para las fuertes densidades, M_1 es evidentemente más densa; en fin, para las densidades medias y grandes, la igualdad puede ser debida o no a una subestimación de M_1 y el resultado es no válido.

- La abundancia de M_1 es menor que la de M_2, el resultado es obviamente válido para las densidades débiles, pero para las densidades medias y grandes, la desviación se puede explicar por la utilización de dos métodos diferentes y el resultado no es válido.

2.4.1. Planificación de las Operaciones

Los dispositivos de medida del muestreo o experimentales involucran siempre puntos ambiguos que necesitan interpretación y decisión. En general se trata de casos extremos y raros, límites borrosos, etc. Para evitar las variaciones en la interpretación y la decisión y por consiguiente para prevenir incoherencias en el sistema de colecta de datos, es preferible enunciar por escrito el plan detallado de operaciones relacionadas con la colecta de datos y agregar toda regla de decisión en la medida en que casos dudosos se presenten. Además, para no omitir ninguna información importante, se preparan fichas para recordar si es necesario las secuencias de las operaciones, unidades de medidas, etc. Esta actividad deberá contener un calendario de las operaciones y si es necesario, cartografiar los puntos o áreas de muestreo que permitirán evitar ciertas contradicciones y poder darse cuenta de la factibilidad del proyecto en términos de adecuación entre los objetivos planteados y los recursos disponibles.

2.4.2. Preprueba

Esta consiste en realizar sobre un pequeño número de elementos un estudio preliminar sujeto a las mismas condiciones y al mismo plan de colecta que el estudio principal. Este estudio piloto juega un papel cuádruple: verificación, colecta de información, optimización y entrenamiento. La **verificación** permite probar la factibilidad global del proyecto, es el desarrollo del método con el fin de identificar las dificultades y los puntos débiles del plan de investigación. La **colecta** de informacion es necesaria previamente, se refiere a la posibilidad de utilizar ciertos planes de muestreo y por otra parte a la necesidad de tener una estimación de la varianza de la población para determinar el número de elementos requeridos para la obtención de una precisión deseada. La **optimización** corresponde esencialmente a la posibilidad

de repartir el esfuerzo de muestreo, ya sea entre los diferentes estratos (muestreo estratificado), entre los diferentes grados de unidades (muestreo por grados), con el fin de maximizar la precisión de las estimaciones para un costo total dado. El **entrenamiento** al personal involucrado en el estudio de iniciar con el método y de definir las ambigüedades, de conocer las manipulaciones, etc. El "pre-test" sirve para mejorar el planteamiento inicial.

2.4.3. Colecta de Datos

Esta fase necesita la previsión de una serie de tácticas para adaptarse a las circunstancias del estudio.

2.4.4. Complejo de Datos

Es el conjunto de datos de resultados en bruto adquiridos durante la colecta de datos, generalmente es presentada en forma matricial, es decir de una tabla con doble entrada donde cada línea de resultados corresponde a las características de un elemento y cada columna a las diferentes variaciones de cada una de las variables.

2.4.5. Tratamiento Informático y Estadístico de Datos

Multiplicar las diferentes técnicas de análisis estadístico respondiendo a la misma necesidad, o buscar diversas maneras de combinar o de transformar los datos hasta que el resultado sea significativo o corresponda a lo que uno está buscando, tiene ciertos riesgos, ya que en lugar de revisar el premodelo conceptual, o de colectar información suplementaria, o de rediseñar un nuevo experimento, se observan en todos los sentidos los datos. Este ejercicio permite casi siempre obtener un resultado significativo en el sentido estadístico; pero la interpretación que surge de estos, frecuentemente es errónea. Es de notarse que mientras las muestras son más grandes, las pruebas detectan más fácilmente pequeñas heterogeneidades cuyo origen puede ser muy diverso incluso los artefactos metodológicos.

2.4.6. Interpretación

Se hace en el marco de los límites del método, en la óptica de la problemática y del premodelo conceptual, en términos de respuesta a las hipótesis de trabajo formuladas al inicio de la investigación y en el espíritu de las pruebas y análisis estadísticos aplicados. Si la metodología ha sido minuciosamente planificada, el número de interpretaciones posibles es único; en el caso inverso aumenta, y la investigación puede entonces perder interés. En una segunda fase, los resultados son confrontados a los resultados de otros autores. Las divergencias y convergencias son interpretadas a la luz de la metodología seguida en los otros estudios.

2.4.7. Conclusiones

Consisten en evidenciar las respuestas seguras proporcionadas por la investigación, a reformular si es necesario la problemática en función de esta aportación, a revisar llegado el caso, el premodelo conceptual, a emitir nuevas hipótesis de trabajo si difieren de las precedentes y en fin, si es esencial, a orientar la metodología de otros estudios. Las conclusiones constituyen así el punto de partida de un nuevo ciclo de procedimiento metodológico.

Ejercicio 1

Estos años surgió una controversia entre el gobierno de Canadá y la Asociación de Protección de la Naturaleza (APN) sobre la evaluación de bebés focas de Groenlandia. La Dirección de Información del Medio Ambiente de Canadá, publicó las informaciones siguientes:

[...] ya que las focas de Groenlandia se concentran para el período de reproducción, esta situación ha permitido hacer una evaluación de la población por conteos aéreos. [...] El punto focal es de localizar diferentes agrupamientos sobrevolando y fotografiándolos por bandas en una dirección y altitud conocidas. A partir de estas fotografías, se hace posible estimar la densidad de focas y obtener un estimado del número total de focas para la región. Como en todos los estudios, existen problemas. El control del vuelo debe ser preciso, si no, la escala de las fotografías no será uniforme. La repartición desigual de los animales sobre el hielo hace difícil la extrapolación, dando por hecho que solo se fotografía una parte. El conteo debe llevarse a cabo antes de que inicie la época de caza y entonces, antes de que las hembras se oculten.

[...] las técnicas actuales de fotografía, han servido al conteo de focas de color oscuro, de mayor edad, que se encuentran sobre el hielo; es difícil distinguir el pelaje blanco de una foca bebé sobre el hielo. Sin embargo, se ha descubierto recientemente que el pelaje blanco de un recién nacido no absorbe las radiaciones ultravioleta del sol, mientras que la nieve sobre la que se encuentra refleja la mayor parte de estas radiaciones. Así, una fotografía con ultravioleta de un recién nacido sobre la nieve, produce una imagen negra. Ya que los bebés nacen en un período de tiempo muy corto y que se quedan casi todos sobre el hielo durante las dos o tres primeras semanas después del nacimiento, debería haber un cierto momento en el que se encuentran casi todos los bebés en el mismo lugar sobre el hielo. Entonces, es teóricamente posible evaluar la producción de bebés de cualquier año, efectuando conteos aéreos y utilizando la fotografía ultravioleta. Se ha ensayado esta técnica llena de promesas y de acuerdo con los resultados del primer conteo, ciertos autores han afirmado que había 95% de oportunidades que la producción de bebés se situara entre 54,683 y 257,602.

Preguntas

Suponiendo que los bebés focas hayan sido contados en un cierto número de corredores de la misma longitud y mismo ancho, repartidos aleatoriamente (muestreo aleatorio simple o sistemático) en la región estudiada:

1. ¿La unidad de muestreo corresponde a un corredor localizado aleatoriamente o a un bebé foca?

2. ¿El número de bebés foca contados en un corredor dado corresponde a los elementos del corredor o a una variante?

3. ¿Los bebés foca recubiertos de nieve no aparecen sobre la fotografía ultravioleta. La subestimación que resulta corresponde a un sesgo del estimador, a un muestreo no representativo de la población estadística (muestra no aleatoria), o a una falta de exactitud del dispositivo de medida o conteo?

4. ¿El número total de bebés focas en la región estudiada corresponde al número de la población estadística o a un parámetro de la misma población?

5. ¿Cuál es la variable estudiada, es cualitativa?

6. ¿La población biológica se confunde con la población estadística?

7. ¿Para que la producción total de bebés focas sea estimada con un mínimo de sesgo, hay que mejorar el dispositivo de medida (o conteo) o aumentar la talla de la muestra?

8. ¿Para que la producción total de bebés focas sea estimada con un máximo de precisión (mínimo de errores aleatorios), hay que aumentar la talla de la muestra o el número de variables a medir?

Ejercicio 2

Los mecanismos que permiten la coexistencia de especies forestales arborescentes son mal conocidos. En efecto, uno se pregunta siempre porque dos especies que comparten la dominancia en las zonas templadas no entran en competencia hasta la eliminación de una de ellas. Varios autores sugieren que el crecimiento de las plántulas de una especie es favorecida bajo la fronda de los adultos de otra especie. Así, tan pronto como una especie predomina, esta crea las condiciones favorables a los individuos jóvenes de su competidor. Para verificar esta explicación, Cypher y Boucher (1982) estudiaron la tasa de crecimiento de plántulas de haya americana con grandes hojas (*Fagus grandifolia ehvih*) bajo la fronda de adultos de haya americana y de maple de azúcar (*Acer saccharum marsh*), por otra parte, también estudiaron el crecimiento de plántulas de maple de azúcar bajo adultos de haya americana y de maple de azúcar. Para minimizar los efectos de diversos factores que pueden afectar al crecimiento, dos sitios fueron escogidos uno a proximidad del otro en el monte San Hilario (Canadá). La fronda del primer sitio está formado por haya americana maduros y la del segundo por maple de azúcar. En cada sitio, 20 plántulas de haya americana y 20 de érable han sido seleccionadas y las tasas de crecimiento han sido medidas tomando la distancia internodular, es decir la distancia entre cicatrices que marcan los fines de crecimiento anual y la formación de un botón terminal.

Preguntas

1. Definir la problemática del estudio.

2. ¿Cuál es la hipótesis de trabajo que los autores se proponen verificar?

3. ¿Esta hipótesis se ajusta al premodelo teórico?

4. ¿Sobre que material biológico trabajan los investigadores?

5. ¿Sobre qué unidad de muestreo (elemento) se trabaja?

6. ¿Cuál (es) población (es) estadísticas estudian y porqué razón?

7. ¿Qué variable cuantitativa se estudia?

8. Cuál es la escala de variación de esta variable si la unidad de medida es el centímetro?

9. ¿La variable a medir es cualitativa, se trata de una variable controlada o de una variable aleatoria que servirá de criterio de clasificación de los elementos de diferentes grupos?

10. ¿Los investigadores han adoptado un enfoque experimental, cuasi-experimental o descriptivo?

Ejercicio 3

En el marco del plan quinquenal de inventario aéreo de jabalí grande de la provincia de Québec, Lachapelle (1980) estimó la densidad de los alces dentro de la zona de caza K_1. El objetivo último del plan era de evaluar el potencial cinegético de diferentes zonas bio-fisico-administrativas de la provincia con el fin de fijar cuotas de caza y sobretodo las fechas de apertura y cierre de la caza, lo cual constituye el medio más cómodo para la administración para controlar la presión de la caza. La zona K_1 que se extiende sobre 22,600 km^2 ha sido subdividida en 375 parcelas de 60 km^2. Con la ayuda de un programa de cómputo de generación de números aleatorios, 20 parcelas han sido seleccionadas y sobrevoladas para el conteo de áreas de confinamiento invernal de grandes mamíferos en el norte de América. Todas las áreas ubicadas en avión han sido revisadas en helicóptero para el conteo e identificación de la edad y el sexo de los individuos. Los resultados se muestran en la tabla opuesta.

Preguntas

1. ¿Cuál es el objetivo último del estudio?

2. ¿Cuál es el objetivo particular del estudio?

3. ¿El objetivo particular corresponde a una hipótesis de trabajo que emana de un premodelo conceptual de explicación?

4. ¿Cuál es la población biológica en estudio?

5. ¿La población biológica se confunde con la población objetivo?

6. ¿Cuál es el elemento (unidad de muestreo) que sirve para estimar la densidad de áreas devastadas ?

7. ¿Cuál es la población estadística que se relaciona con la estimación de la densidad de las áreas (parcelas)?

8. ¿La población estadística se confunde con la población biológica?

9. ¿La población estadística es finita, si así es, cual es el número total de elementos?

10. ¿La muestra extraída de esta población estadística es una muestra aleatoria simple?

11. ¿Qué variables se miden en cada elemento?

12. ¿Cuál es la escala de variación de cada variable?

13. ¿Si se desea estimar el porcentaje de machos en adultos o en juveniles, se puede considerar al alce como una unidad de muestreo?

14. ¿Si así es, cual es el plan de muestreo y llegado el caso, definir las diferentes unidades de muestreo?

15. ¿La población estadística ha cambiado de naturaleza?

16. ¿Si es así, se conoce el número total de elementos?

17. ¿El enfoque seguido en este estudio es experimental o descriptivo?

Número de la parcela	Número del área	Número de machos adultos	Número de hembras adultas	Número de alces
UK 20 W	1	0	0	0
QE 05 E	1	0	1	0
PE 42 E	1	0	1	0
PE 93 E	1	3	0	0
	2	1	0	0
	3	0	1	0
	4	0	2	0
	5	0	1	0
PE 28 W1	1	1	0	0
	2	1	0	0
PE 69 E	0	0	0	0
PE 86 W	0	0	0	0
PE 72 E	1	1	2	0
	2	0	1	1
	3	0	1	2
	4	0	2	0
PE 16 E	1	0	2	0
	2	0	0	0
PE 56 W	1	1	1	0
	2	0	1	1
	3	0	1	1
PE 85 E	1	1	1	1
	2	1	0	0
	3	0	1	1
PE 64 W	1	1	0	0
	2	0	1	0
UJ 27 W	1	1	2	1
	2	0	0	0
TK 82 W	1	0	2	2
	2	1	0	0
	3	0	1	0
	4	0	1	0
	5	1	0	0
	6	4	0	0
TK 95 E	0	0	0	0
UK 06 W	1	0	2	2
UK 07 E	1	1	3	0
	2	0	1	1
PD 99 W	1	1	1	2
	2	0	0	0
	3	0	2	0
UK 01 E	1	1	1	2
	2	0	1	0
	3	1	1	1
	4	0	2	3
UK 00 E	1	1	1	2

Ejercicio 4

Sphaerophoria philantus mg es un insecto Syrphidae que en estado larvario es un conocido predador de los pulgones. Sin embargo, su impacto es considerado como real en la medida en que los huevos de este Syrphidae son puestos en la proximidad de colonias de pulgones.

Para verificar si la hembra de *S. philantus* deposita sus huevos en función de la presencia de pulgones, 30 plantas de maíz han sido seleccionadas aleatoriamente en el seno de una parcela experimental. Sobre cada planta así seleccionada, los pulgones y los huevos de Syrphidae han sido contados durante el período de puesta del depredador. Los resultados de estos conteos son los siguientes:

No de planta	Número de pulgones	Número de huevos	No de planta	Número de pulgones	Número de huevos
1	25	1	16	45	2
2	10	1	17	55	2
3	75	3	18	0	0
4	250	6	19	125	4
5	10	1	20	35	1
6	50	2	21	10	0
7	100	3	22	75	3
8	85	2	23	50	2
9	0	0	24	0	0
10	0	0	25	30	1
11	65	2	26	20	1
12	30	1	27	200	5
13	5	0	28	40	1
14	400	8	29	70	3
15	35	1	30	15	1

Preguntas

1. ¿Qué poblaciones biológicas se estudian?

2. ¿Cuál es la población estadística, está definida?

3. ¿Cuál es el elemento?

4. ¿El estudio es realizado siguiendo un enfoque descriptivo, experimental o cuasi-experimental?

5. ¿La muestra es representativa de la población?

6. ¿Cuál es el número de elementos de la muestra?

7. ¿Si uno aumenta la talla de muestra, su representatividad se aumenta?

8. ¿Si se aumenta la talla de la muestra, la estimación de diferentes parámetros de la población será más precisa?

9. ¿Cuál es la hipótesis de trabajo?

10. ¿Cuáles son las variables medidas?

11. ¿Son las variables pertinentes?

12. ¿Son estas variables aleatorias?

13. ¿Qué factores podrían afectar la exactitud y la fidelidad de los conteos?

14. ¿Si se procede a un doble conteo de pulgones y de huevos, se mejora la fidelidad o la exactitud?

15. ¿Si el número de pulgones fuera sistemáticamente subestimada, un aumento de la talla de muestreo podría compensar este error?

2.5. Distribución de Frecuencias (Arreglo ordenado)

Literatura sugerida:

Daniel.W. W., 1982, Biostadística, Limusa. México. 485 p (pág. 15-16)

King, B. M., Minium E. M., 2003, Statistical Reasoning, 4a edición. Estados Unidos. 550 p (pág. 52-53).

Cuando los datos estadísticos de los cuales se dispone son numerosos, el trabajar con ellos directamente es complicado y poco se puede hacer con ellos si nos se les organiza y clasifica, es decir se les arregla de acuerdo a algún método y se plasman en una tabla de recuento. Este método estadístico es conocido como **distribución de frecuencias**.

Para comprender la técnica de distribución de frecuencias y dominar sus aplicaciones, es preciso saber que es un **intervalo de clase**. Los intervalos de clase están limitados por valores extremos que se denominan **límite inferior (Li)** y **límite superior (Ls)**, por ejemplo:

Intervalo de clase 22 - 25 (Clase de valores de 22 a 25)

Intervalo de clase 25 - 27 (Clase de valores de 25 a 27)

En este ejemplo 22 es el límite inferior de la clase 22 a 25, pero un valor exactamente de 25 ¿A qué clase pertenece? Para evitar esta ambigüedad pudiera tomarse la clase de 22-25 y 26-27, ahora no hay ambigüedad pero hemos dejado sin pertenecer a ninguna clase los valores intermedios de 25 a 26. Esto conduce al concepto de **límites reales de clases** que corresponden al punto medio del límite superior de una clase y el límite inferior de la siguiente. Así, los límites de las clases anteriores serían:

21.5 – 25.5

25.5 – 27.5

Si la variable no toma ninguno de estos valores intermedios el problema está resuelto, pero si la variable tomara el valor 25.5 entonces las clases pudieran escribirse:

21.5 – 24.49

24.5 – 27.49 etc

Anchura o tamaño de un intervalo de clase. Es la diferencia entre los límites de clase.

Marca de clase, es el valor correspondiente al punto medio de un intervalo de clase y su valor es igual a la mitad de la suma de los límites de clase.

La organización de los datos en una distribución de frecuencias permite estudiar su comportamiento, y consiste en arreglar los datos ordenándolos en intervalos de clase e indicando el número de datos comprendidos en cada clase.

En toda serie de datos estadísticos existen valores extremos y la diferencia entre ellos es utilizada para definir el número de intervalos de clase.

Como determinar el **número de clases** o **intervalos de clase**; existen dos reglas principales. La regla de Sturge y la regla de Yule.

$K = 1 + 3.32(\text{Log}_{10} n)$ *regla de Sturge*

Donde: ***n***= numero de elementos disponibles en la muestra

No. Clases o intervalos de clase (K) $= 2.5 * 4\sqrt{n}$ *regla de Yule*

Donde: ***n***= numero de elementos disponibles en la muestra

El número de **intervalos de clase** depende de la distribución que quiera hacerse, si son muy pocos, se pierden detalles y si son muchos, se manifiestan irregularidades que no permiten apreciar un patrón de comportamiento. En todo caso la mayoría de analistas recomienda no menos de 5 ni más de 15 intervalos de clase.

Las reglas generales para formar distribuciones de frecuencia son las siguientes:

1. Se obtiene la diferencia entre los valores extremos (Amplitud, R).

   R= dato mayor – dato menor

2. Se calcula el número de intervalos de clase (ecuación anterior)

   $K = 1 + 3.32(\text{Log}_{10} n)$

3. Se estima el tamaño de cada intervalo (W) mediante:

$$W = \frac{R}{K}$$

Si el resultado de la división no es un número entero, se recomienda “redondear” al entero superior. Así si la división dio 3.5 el número de intervalos se toma como 4.

Nueva amplitud (NR) = (W)(K)

Si la amplitud fue de 21 y el número de intervalos de clase que se calculó de 6.

W = 21/6 = 3.5
W = 4 por lo cual,
La nueva amplitud será 6(4) = 24

El exceso de 3 que se tiene con relación al rango original, se distribuye entre el límite superior y el limite inferior; al agregar 2 al limite superior y restar 1 al inferior o viceversa. De ambas formas la diferencia entre los valores extremos es de 24.

Para la selección del número de intervalos de clase no pueden darse reglas invariables; el número se selecciona atendiendo a diversos factores tales como el rango, variabilidad de los datos e incluso finalidad del estudio estadístico.

4. Se forman los intervalos de clase agregando (K) al límite inferior de cada clase, principiando por el número inferior de la diferencia entre los valores extremos.

5. Se fijan los límites reales de cada clase.

6. Se determinan las frecuencias de cada clase.

Ejemplo 1: Un investigador desea determinar como varían las tallas de mojarras de una laguna tomando una muestra de 50 organismos y anota sus tallas en mm, encontrando:

65	53	64	60	68	63	57	63	61	62
63	58	72	92	66	65	59	69	55	60
68	64	56	63	61	67	63	61	66	65
59	70	62	66	69	64	55	64	61	67
64	58	67	65	57	71	62	66	64	60

Se elabora un cuadro de distribución de frecuencias de la siguiente manera:

1. Obtención de las diferencias entre los valores extremos, amplitud o rango es = 72 – 53 = 19

2. Se obtiene K = 7 intervalos de clase ($K = 1 + 3.32(\text{Log}_{10}n = 6.64 \approx 7$).

3. Para de terminar la talla del intervalo de clase tenemos:

$W = 19/7 = 2.71 \approx 3$

$W = 3$

La nueva amplitud = 3 (7) = 21; el exceso de dos se distribuye de la siguiente manera:

53 – 1 = 52 y 72 + 1 = 73

4. Formamos los intervalos agregando al límite inferior o sea 52 + 3 = 55 que será el límite superior de la primera clase.

5. Encontramos los límites reales agregando 0.5 a los límites de cada clase, es decir:

52.5 a 55.5

6. Se cuentan las frecuencias que caen en cada intervalo, a tarves del recuento.

7. Se calcula la marca d clase, sumando los límites superior e inferior y dividiéndolas entre 2, ya que la marca de clase es el valor central de cada intervalo de clase.

8. Se calcula la frecuencia relativa, para obtener el valor relativo que representa cada frecuencia con respecto al total,

9. Se calcula las frecuencias acumuladas y las frecuencias relativas acumuladas.

10. Toda la información se inserta en una tabla de recuento y frecuencia de las tallas para las 50 mojarras de la siguiente manera:

Intervalos de clase (mm)	**Marcas de clase (xi)**	**Recuento**	**Frecuencia *f(xi)***	**Frecuencia relativa *fr(xi)***	**Frecuencia acumulada *fc(xi)***	**Frecuencia acumulada relativa *frc(xi)***
52.5 – 55.5	54	///	3	0.06	3	0.06
55.5 – 58.5	57	𝍸	5	0.10	8	0.16
58.5 – 61.5	60	𝍸 𝍸	9	0.18	17	0.34
61.5 – 64.5	63	𝍸 𝍸 𝍸	15	0.30	32	0.64
64.5 – 67.5	66	𝍸 𝍸 /	11	0.22	43	0.86
67.5 – 70.5	69	𝍸	5	0.10	48	0.96
70.5 – 73.5	72	//	2	0.04	50	1.00

2.5.1 El Histograma

Es la representación gráfica de la distribución de frecuencias, se pueden utilizar en variables cualitativas y cuantitativas. En el caso de una variable cuantitativa, para elaborar un histograma primero se señalan en la horizontal las marcas de clase y luego para cada marca de clase se señala una barra cuya altura será la frecuencia absoluta o relativa correspondiente (Figura 6). En la tabla de recuento $f(x_i)$ significa frecuencia absoluta de clase; $fr(x_i)$ frecuencia relativa.

La frecuencia relativa es la proporción que ocupa cada valor de frecuencia absoluta con respecto al tamaño de la muestra ***(n)***; y se obtiene dividiendo la frecuencia de cada clase por ***n***:

$$fr\,(x_i) = \frac{f\,(x_i)}{n}$$

La frecuencia acumulada $fc(x_i)$ es el valor acumulado de las frecuencias reales a cada marca de clase en donde el último valor de esta columna será igual a ***n***. De manera análoga, $fcr(x_i)$ designa la variable acumulada relativa de cada clase. Por ejemplo: ***frc(63)*** es igual a 0.64 porque la proporción de veces en que ***x*** toma el valor de 63 o menos es 0.64, es decir que el 64% de las mojarras tienen una talla de 63 mm o menos:

$$frc\,(x) = \frac{fc(x_i)}{n}$$

Las frecuencias acumuladas relativas de la última columna se pueden obtener sumando sucesivamente las frecuencias relativas.

El polígono de frecuencia es la unión a través de líneas rectas de los puntos medios de la parte superior de cada barra del histograma (figura 6).

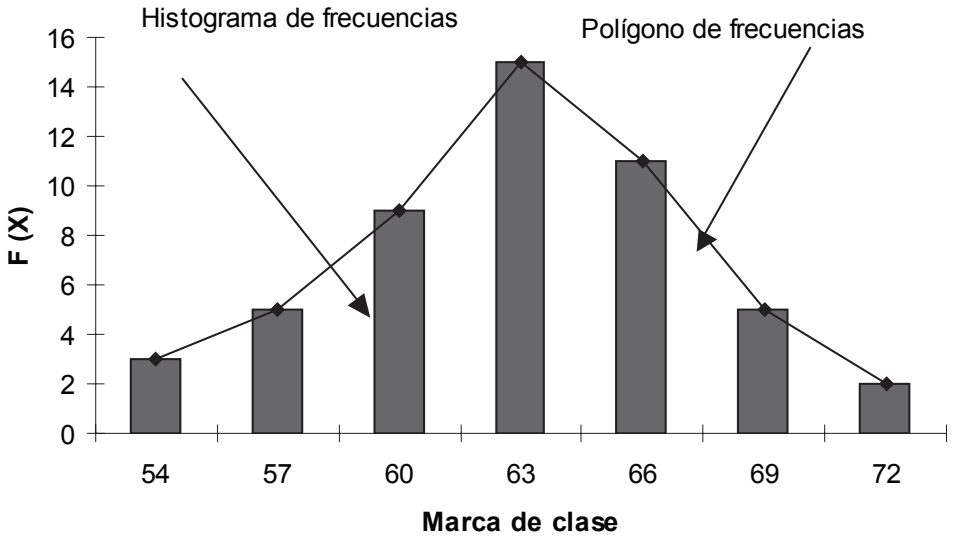


Figura 6. Histograma correspondiente a los datos del ejemplo 1

2.5.2. Diagramas de Barras

En el caso de una variable cualitativa, se utilizan los diagramas de barras y son una representación de la frecuencia de la variable estudiada y se expresan en frecuencias relativas; así por ejemplo se puede dibujar en un diagrama de barras la frecuencia relativa del número de empleos que se registraron de los profesionistas con doctorado en diferentes instituciones (figura 7).

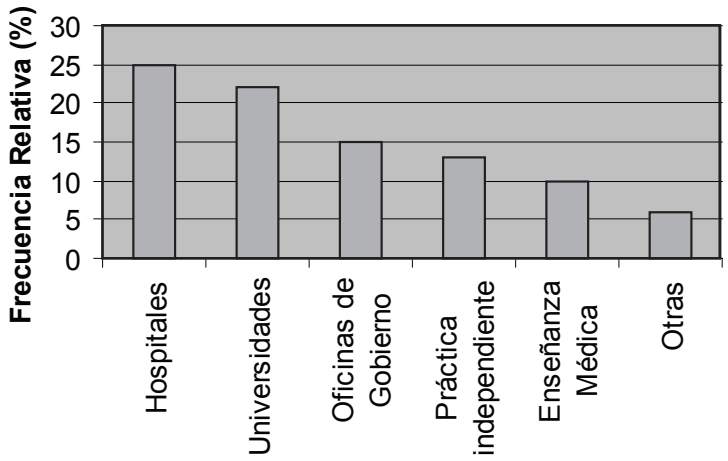


Figura 7. Diagrama de barras de una variable cualitativa (modificado de King, y, Minium (2003), pág 83)

2.5.3. Ojivas o Curvas Sigmoides

La distribución de frecuencia relativa acumulada, permite realizar curvas de porcentaje acumulado, llamadas **"ojivas"** o ***curvas sigmoides***, este tipo de curvas permite describir a la variable estudiada en función de la representación de la frecuencia en términos de porcentaje, es sobretodo común para describir el valor de la variable que agrupa al 50% de los datos, al 25% y al 75%, de esta manera, los datos descritos en la tabla anterior tendrían la siguiente curva (figura 8):

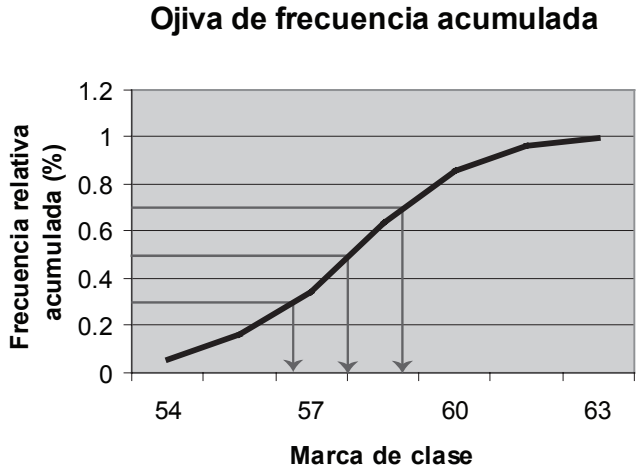


Figura 8. Ojiva de distribución de frecuencias relativas

2.5.4. Gráficas de Pay

Otra representación común de la información, son las gráficas de pay, se utilizan generalmente para representar variables cualitativas con frecuencias relativas. De esta manera cualquier área del pay, es una fracción de la frecuencia del número total de casos en la distribución. En el caso del ejercicio del diagrama de barras anterior, su representación sería (figura 9):

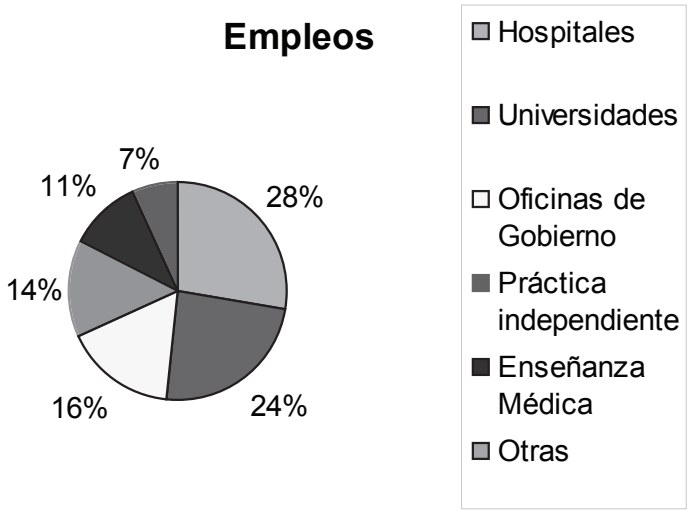


Figura 9. Representación en pay de una variable cualitativa (modificado de King y Minium (2003), p. 83)

2.6. Estimadores (estadígrafos y parámetros)

Los parámetros son las medidas estadísticas que representan a los elementos de una población y que pueden ser la media, desviación estándar, varianza, número de elementos, etc. De manera análoga las medidas que representan a los elementos de la muestra tienen la misma denominación, sin embargo se diferencian por su simbología y se les denomina estadígrafos:

Literatura sugerida:

Scherrer B., 1984, Biostatistique. Gaëtan Morin Editor. Montreal, Paris, Casa Blanca. 850 p. (Pág 132-135).

Zar. J. H., 1999. Bostatistical Analysis (4 edición). Prentice Hall. Estados Unidos. 663 p. (Pag. 20-29).

Daniel.W. W. 1982, Biostadística. Limusa. Mexico. 485 p. (pág. 6-11)

Medida	Parámetro- Población	Estadígrafo-Muestra
Media aritmética	μ	x
Proporción	*P*	*p*
Varianza	σ^2	s^2
Desviación estándar	σ	S
Número de elementos	*N*	n

2.6.1 Medidas de tendencia central

Al estudiar la información estadística mediante los histogramas y los polígonos de frecuencia, se observó el comportamiento de los datos en cuanto a la frecuencia con que se presentan los valores; siendo algunos más frecuentes que otros. Además se observó una tendencia de agrupación alrededor de los datos más frecuentes, haciendo que la distribución de frecuencias y/o histograma adquiera una forma de campana. Por lo general, la mayor densidad de frecuencia está en la parte central de las gráficas, y es de aquí donde se deriva el concepto de medidas de tendencia central, que es el nombre que recibe la **media, mediana** y **moda**.

2.6.1.1 La media aritmética

Suele simplemente llamarse **"la media"** o **"promedio",** es la medida de tendencia central más utilizada. La media para datos no agrupados se obtiene a través de la suma de los valores observados dividida por el número total de observaciones, designándose la media de la población y de la muestra por:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \ldots + x_n}{N} \qquad \bar{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

Media poblacional — Media muestral

Donde: ***N*** y ***n*** representan el número de elementos de la población y de la muestra respectivamente, ***x*** el valor de cada elemento, y ***μ y*** $\bar{X}$ la media de la población.

La **media ponderada** o **media de una distribución de frecuencias**, se calcula a partir de una distribución de frecuencias, tomándose cada marca de clase como el valor representativo de ese intervalo de clase. Cada marca de clase es multiplicada por su respectiva frecuencia, sumándose los productos y dicha suma se divide entonces por el número total de observaciones, describiéndose a través de la siguiente fórmula:

$$\mu = \frac{\sum_{i=1}^{N} x_i\, fr(x_i)}{N} = \frac{x_1 fr(x_1) + x_2 fr(x_3) + \ldots + x_n fr(x_n)}{N} \qquad \textit{media poblacional}$$

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i\, fr(x_i)}{n} = \frac{x_1 fr(x_1) + x_2 fr(x_3) + \ldots + x_n fr(x_n)}{n} \qquad \textit{media muestral}$$

Así partiendo del ejercicio anterior la media calculada sería 62.94 mm

Al calcular la media aritmética con datos agrupados, su valor se aproximará al valor obtenido con datos no agrupados. **El valor de la media no será representativo de la población si la distribución de frecuencias es muy irregular, demasiado asimétrica o sesgada**.

Ya que la media es una medida de tendencia central, debería en todas las ocasiones indicar el centro de una serie de observaciones. Una media aritmética es muy sensible a la influencia de unos cuantos números grandes en un conjunto de pocas observaciones; en estos casos la media no representa el centro de una serie de valores.

Suponga que usted coloca 1000 trampas para la captura de cierta especie en su medio silvestre. De las trampas colocadas, 100 atraparon cada una 200 organismos y las 900 trampas restantes cero organismos, provocando que el 90% de las trampas con valores de cero minimicen por completo el "peso" del 10% restante. En algunos casos tal sensibilidad de la media es a veces deseable, sin embrago en algunos casos no es así, y es entonces cuando es necesario utilizar otras medidas de tendencia como la *mediana*.

Ventajas de la media:

- Fácil de calcular
- La suma de los cuadrados de las desviaciones de la media es poca o igual a la suma de los cuadrados de las desviaciones de la mediana y la moda.

Inconvenientes de la media:

- Fuertemente influenciada por sus valores extremos
- Representa muy mal los valores de una población heterogénea (distribución polimodal o fuertemente asimétrica.

Existen otros dos tipos de media, la media geométrica y la media armónica.

2.6.1.2 La media geométrica

Se define como la raíz (**n**) del producto de **n** términos, así si los datos son $\mathbf{x_1, x_2, \ldots, x_n}$ su media geométrica es:

$$Media\_Geométrica = \sqrt[n]{x_1 x_2 x_3 \ldots x_n} = n\sqrt{\prod_{i=1}^{n} x_i} = antilog \frac{\sum_{i=1}^{n} log X_i}{n}$$

Por ejemplo la media geométrica de 2, 4, 6, 12, 18 es:

$$\sqrt[5]{(2)(4)(6)(12)(18)} = antilog\left(\frac{log(2) + log(4) + log(6) + log(12) + log(18)}{5}\right) = 6.36$$

La media geométrica es útil en el cálculo de tasas de crecimiento: Así por ejemplo, si el crecimiento de las ventas de un negocio fue en los últimos tres años de 26%, 32%, 28%, hallar la media anual de crecimiento:

$$\sqrt[5]{(1.26)\,(1.32)\,(1.28)} = antilog\left(\frac{log\,(0.26) + log\,(0.32) + log(0.28)}{3}\right) = 1.286$$

∴ El crecimiento fue de 28.6%

2.6.1.3. La media armónica (H)

La media armónica de una serie de números es el recíproco de la media aritmética de los recíprocos de los números de la serie. Sean los números x_1, x_2, x_3,..., x_n; la media armónica ***H*** se obtendrá de la relación:

$$Media\ Armónica\ H = \frac{n}{\frac{1}{x_1}+\frac{1}{x_2}+\frac{1}{x_3}+....+\frac{1}{x_n}}$$

Por ejemplo: Calcular la media armónica de los números 4, 5, y 8

$$H = \frac{3}{\frac{1}{4}+\frac{1}{5}+\frac{1}{8}} = \frac{3}{\frac{23}{40}} = \frac{23}{120} = 5.22$$

La media armónica es ocasionalmente utilizada cuando se trata con tasas promedio de superficie. Si todos los datos son idénticos entonces la media armónica es la misma que la media aritmética, así como también a la media geométrica; y si una serie es positiva pero no idéntica, entonces la media armónica < media geométrica < media aritmética.

2.6.2. La Mediana

Se define como el valor que divide una distribución de datos ordenados en dos mitades, o sea la medida que deja por arriba igual número de términos que por debajo de él. En otras palabras **la mediana es el valor del término del punto medio de una serie de valores**.

Para el cálculo de la media aritmética no importa si los datos estén o no ordenados, a diferencia de la mediana que **requiere para su cálculo, que el conjunto de valores se encuentren ordenados de menor a mayor**.

La mediana para datos no agrupados está dada por:

$$Med = x_{(n+1)} / 2$$

Donde x corresponde a la variable y n al número de datos

Por ejemplo: Se tiene la siguiente serie de observaciones:

Datos	Valor
x_1	0
x_2	1
x_3	1
x_4	2
x_5	3
x_6	**4**
x_7	6
x_8	7
x_9	8
x_{10}	8
x_{11}	9

Es decir que la mediana esta localizada en el sexto dato que es igual a 4.

Para encontrar la **mediana de datos agrupados** se suele utilizar el método de interpolación; que requiere construir una tabla de distribución de frecuencia acumulada. Se basa en el supuesto de que las observaciones de cada clase estén igualmente repartidas en los intervalos, la cual se describe como sigue:

$$Med = L_M + \frac{i}{fr(x_i)}\left(\frac{n}{2} - fc(x_{i-1})\right)$$

Donde: $fc(x_{i-1})$ es la frecuencia acumulada bajo el límite inferior (anterior) de la clase mediana; $\mathbf{fr(x_i)}$ es la frecuencia de la clase mediana; $\mathbf{L_M}$ es el límite inferior de la clase mediana e ***i*** es el tamaño del intervalo de clase. Tomando la tabla de distribución de frecuencias del ejemplo 1, tenemos:

Intervalos de clase (mm)	Marcas de clase (x_i)	Recuento	Frecuencia $f(x_i)$	Frecuencia relativa $fr(x_i)$	Frecuencia acumulada $fc(x_i)$	Frecuencia acumulada relativa $frc(x_i)$
52.5 – 55.5	54	///	3	0.06	3	0.06
55.5 – 58.5	57	////	5	0.10	8	0.16
58.5 – 61.5	60	//// ////	9	0.18	17	0.34
61.5 – 64.5	63	//// //// ////	15	0.30	32	0.64
64.5 – 67.5	66	//// //// /	11	0.22	43	0.86
67.5 – 70.5	69	////	5	0.10	48	0.96
70.5 – 73.5	72	//	2	0.04	50	1.00

$$Med = 61.5 + \frac{3}{15}\left(\frac{50}{2} - 17\right) = 63.1$$

Ventajas de la mediana:

- Puede ser calculada para valores de características cíclicas (estaciones, horas) para las que la media tiene poco significado.
- No esta influenciada por valores extremos

Inconvenientes de la mediana:

- Se presta mal a cálculos estadísticos o algebráicos
- No representa que el valor que separa la muestra en dos partes iguales sin tener en cuenta el conjunto de datos.

2.6.3. La Moda

Puede ser definida como el valor que ocurre con mayor frecuencia en una distribución. Para calcular el valor de la moda en una distribución de frecuencias, es necesario primero identificar la clase en la cual se sitúa la moda, la cual corresponde a aquella que presenta la máxima frecuencia de la distribución.

Así en el ejemplo 1 de las mojarras, la clase 61.5 a 64.5 es la clase modal por tener la mayor frecuencia. Una vez identificada la clase modal, el paso siguiente es identificar la moda dentro de la clase.

Al aproximar un valor modal por interpolación, hay que tomar en cuenta que los datos, generalmente, tienden a agregarse hacia el punto de mayor densidad. Hay una tendencia en este punto de concentración a llevar el valor modal ya sea hacia el límite superior o hacia el límite inferior de la clase, según que la clase sea la post-modal o la pre-modal, entonces para el cálculo aproximado de la moda tenemos:

$$Mo = \left(\frac{d_i}{d_i + d_s}\right) i + L$$

Donde: $\boldsymbol{d_i}$ es la diferencia absoluta entre la frecuencia de la clase modal y la clase inferior próxima; $\boldsymbol{d_s}$ es la diferencia absoluta entre la frecuencia de la clase modal y clase superior próxima; $\boldsymbol{i}$ es tamaño del intervalo de clase; y $\boldsymbol{L}$ es el límite inferior de la clase modal. Así, la posición de la moda puede ser una medida útil de la tendencia de la variable a estudiar, sin embargo es considerada la medida de tendencia central más imprecisa.

En el ejemplo anterior la moda se calcula como sigue:

$$Mo = \left(\frac{(15\text{-}9)}{(15\text{-}9) + (15 - 11)}\right) 3 + 61.5 = 63.3$$

Por lo que la moda es 63.3.

Ventajas de la moda:

- No esta afectada por sus valores extremos
- Puede ser calculada en variables cíclicas
- Es buen indicador de poblaciones heterogéneas

Inconvenientes de la moda:

- Puede variar si se modifica el intervalo de clase

2.7. Medidas de Dispersión

Literatura sugerida:

Scherrer B., 1984, Biostatistique. Gaëtan Morin Editor. Montreal, Paris, Casa Blanca. 850 p (Pág 157-168).

Zar. J. H., 1999. Biostatistical Analysis. 4 edición. Prentice Hall. Estados Unidos. 663 p. (Pag. 32-40).

Daniel.W. W. 1982, Biostadística, Limusa. Mexico. 485 p (pág. 11-14)

Las medidas de dispersión nos indican la posible desviación de los datos, respecto a la medida de tendencia central "la media". Las medidas de tendencia central no describen el comportamiento de los datos con relación a como estos se dispersan en una distribución de frecuencias alrededor de tendencia central; además poco nos indican sobre un determinado dato con relación a los demás en una distribución. Entre las medidas de dispersión tenemos a la **amplitud o rango, la desviación estándar, la varianza, el coeficiente de variación, cuantiles y la desviación media**. En esta antología solo mencionaremos a los más utilizados que son las cuatro primeras.

2.7.1. Amplitud

En toda distribución los límites inferior y superior determinan el "recorrido" de los datos, por lo cual la **amplitud** que **es la diferencia entre los límites superior e inferior**, de una distribución es una mediada de dispersión de los datos. Como se observa esta medida es la más fácil de obtener, sin embargo es poco utilizada debido a que no denota la presencia de valores extremos de poca frecuencia o los valores "de alto ruido". Para **datos agrupados** se considera a **la amplitud** como l**a diferencia entre el límite superior de la clase más alta y el límite inferior de la clase más baja**.

2.7.2. La Varianza y Desviación Estándar

La varianza indica el promedio de la suma de las desviaciones cuadráticas, en torno al valor promedio de la variable. La varianza de la muestra s^2 es la suma de los cuadrados de las desviaciones respecto a la media aritmética.

Para datos no agrupados es como sigue:

$$S^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{X})^2}{n - 1}$$

Muestra

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

Población

Para datos agrupados el procedimiento es mediante la siguiente ecuación:

$$S^2 = \frac{\sum_{i=1}^{n} fx_i (x_i - \bar{X})^2}{n - 1}$$

Muestra

$$\sigma^2 = \frac{\sum_{i=1}^{N} fx_i (x_i - \mu)^2}{N}$$

Población

Como ejemplo de aplicación podemos utilizar los datos siguientes: supóngase que el valor promedio de longitudes de juveniles de camarón de una muestra es de 9 cm, los datos de la variable (x_i) se deberán restar a cada valor de la media (a esto se le llama desviación media), después sumarlos, elevarlos al cuadrado y dividirlos entre el numero total de elementos (***n***), asi el valor de la varianza calculada es:

$$S^2 = \frac{(10-9)^2 + (12-9)^2 + (2-9)^2 + (9-9)^2 + (15-9)^2 + (6-9)^2 + (7-9)^2 + (8-9)^2 + (12-9)^2 + (9-9)^2}{10} = 11.8$$

$$S^2 = \frac{(10)^2 + (12)^2 + (2)^2 + (9)^2 + (15)^2 + (6)^2 + (7)^2 + (8)^2 + (12)^2 + (9)^2}{10} = 11.8$$

La varianza se expresa en unidades cuadradas y no es la unidad original debido a la operación de elevar al cuadrado que se ha efectuado. Por consiguiente es preciso obtener la raíz cuadrada para volver a la unidad original. La medida de dispersión que así se obtiene se llama **desviación típica o estándar s**; la cual se expresa de la siguiente forma:

$$\sigma^2 = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$$

Para datos no agrupados (Población)

$$S = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \overline{X})^2}{n-1}}$$

Si se trata de la muestra

$$S = \sqrt{\frac{\sum_{i=1}^{n} fi(x_i - \overline{X})^2}{n-1}}$$

Para datos agrupados

Con el ejemplo señalado previamente de la serie con varianza igual a 11.8 cm, la desviación estándar se puede calcular como la raíz cuadrada de la varianza cuyo resultado es de 3.43 cm.

2.7.3. El Coeficiente de Variación

La desviación estándar es útil como medida de la variación dentro de un conjunto dado de datos. Sin embargo, cuando se desea comparar la dispersión de dos conjuntos de datos, comparar las desviaciones estándar puede conducir a resultados ilógicos. Puede ser que las dos variables que intervienen se midan en unidades diferentes. Por ejemplo, es posible que se desee saber, para cierta población, si los niveles de colesterol en el suero, medidos en mg por 100 ml, son más variables que el peso del cuerpo,

medido en kilogramos. Es más aunque se use la misma unidad de medición, las dos medidas pueden ser bastante diferentes. Si se compara la desviación estándar de los pesos de los niños de primer año de primaria, con la desviación de los pesos de los jóvenes de primer año de secundaria, es posible que se encuentre que las desviaciones estándar de estos últimos es numéricamente mayor que la de los primeros, porque los propios pesos son mayores, no porque la dispersión sea mayor. Lo que se necesita en situaciones como esta, es una medida de variación relativa, en lugar de una variación absoluta. Esta medida se encuentra en el coeficiente de variación, el cual expresa a la desviación estándar como un porcentaje de la medida.

$$CV = \frac{S}{\overline{X}} * 100$$

Se observa que como la media y la desviación estándar se expresan en la misma unidad de medición, esta unidad se cancela al calcular el coeficiente de variación, entonces lo que se tiene es una medida independiente de la unidad de medición, y ahora se expresa en porcentaje.

2.8. Probabilidad

Literatura sugerida:

Daniel.W. W. 1982, Biostadística. Limusa. Mexico. 485 p (pág. 33)

Scherrer B., 1984, Biostatistique. Gaëtan Morin Editor. Montreal, Paris, Casa Blanca. 850 p (Pág 195).

Zar. J. H., 1999. Bostatistical Analysis (4 edición). Prentice Hall. Estados Unidos. 663 p. (Pag. 49-62).

http://facultad.sagrado.edu/ConceptosBasicos.pdf#search=%22tabla%20de%20probabilidad%20conjunta%22

2.8.1 Introducción a la Probabilidad

La teoría de la probabilidad proporciona la base para la inferencia estadística. El concepto de probabilidad no es extraño para los profesionales que trabajan en las ciencias naturales, y con frecuencia forma parte de nuestra conversación cotidiana. Por ejemplo es común escuchar decir a un medico que un paciente tiene un 50% de probabilidad de sobrevivir a cierta enfermedad al igual que un agricultor, ganadero o acuacultor hablar sobre probabilidad de sobrevivencia.

Aunque los primeros trabajos sobre probabilidad fueron realizados por Giralamo Cardano (1501-1576) y Galileo Galilei (1564-1642), la investigación formal de la probabilidad como una rama de las matemáticas surgió formalmente desde 1654 entre dos grandes matemáticos franceses, Blaise Pascal (1623-1662) y Pierre de Fermat (1601-1665). Estos dos hombres estuvieron motivados por el deseo de predecir los resultados en los juegos de azar populares dentro de la nobleza francesa durante el siglo XVII.

2.8.2 Probabilidad de un Evento

El cálculo de **los posibles resultados de un "evento**", supone que este puede ocurrir de ***k*** maneras, pero en solo una de esas maneras a la vez. Por ejemplo una moneda tiene dos lados y cuando se arroja al suelo puede caer con "águila" (H) o "sol" (T) hacia arriba, pero solo uno de los dos lados a la vez. Otro ejemplo es en relación a un dado. Este tiene seis lados también al arrojarlo puede tomar un solo valor de

las seis opciones que presenta. En este contexto la probabilidad va a referir **a los posibles resultados** (H o T con la moneda; 1, 2, 3, 4, 5, 6, con el dado) como un **evento**.

Si algún fenómeno puede ocurrir en alguna de las k_1 formas diferentes y también en alguna de la k_2 formas diferentes, entonces el número posible de formas para que ambos eventos ocurran es $k_1 \times k_2$. Por ejemplo, si dos monedas son lanzadas, entonces existirán dos posibles resultados en una de las dos monedas (H, T) y dos en la otra (H, T). Por lo tanto $k_1=2$ y $k_2=2$, donde existen $(k_1)(k_2) = (2)(2) = 4$ posibles resultados de los lanzamientos de ambas monedas: (a) Ambas sol; (b) sol y águila; (c) Águila y sol (d) Ambas águila. *e.g.* (H, T; T, H; H, H; T, T).

Podemos también considerar el lanzamiento de una moneda y un dado; donde habrá dos posibles resultados para la moneda ($k_1=2$) y seis posibles resultados para el dado ($k_2 = 6$); luego entonces tendremos en total (2)(6) = 12 posibles resultados de los dos fenómenos juntos:

H,1; H,2; H,3; H,4; H,5; H,6; T,1; T,2; T,3; T,4; T,5; T,6

Entonces la regla mencionada arriba se extiende fácilmente para determinar el número de maneras (k) en que n elementos pueden ocurrir juntos. Si un objeto de estudio puede ocurrir de k_1 maneras y un segundo objeto de estudio en k_2 maneras, un tercero en k_3 y así sucesivamente, a través de ***n-ésimo*** objeto de estudio y las k_n formas; entonces el número de maneras para que todas las n cosas ocurran juntas es:

$$(k_1)\,(k_2)\,(k_3)\ldots k_n \qquad o \qquad \prod_{i=1}^{n} k_i$$

Ejemplo 2. Un arreglo lineal de tres nucleótidos de ADN es denominado como un triplete. Un nucleótido quizá contenga alguna de las cuatro bases nitrogenadas: Adenina (A), Citosina (C), Guanina (G) y Timina (T). ¿Cuántos diferentes tripletes son posibles?

Como el primer nucleótido en el triplete puede presentar una de las cuatro bases nitrogenadas (A, C, G, T) y el segundo nucleótido también puede presentar una de esas cuatro bases, al igual que el tercero; entonces los posibles resultados de un triplete serán:

$(k_1)(k_2)(k_3) = (4)(4)(4) = 4^3 = 64$ posibles resultados

Ejemplo 3. Si una célula diploide contiene tres pares de cromosomas y solo un miembro de cada par se encuentra en cada gameto, ¿Cuántos gametos diferentes son posibles?

Como el primer cromosoma puede ocurrir en un gameto en una de sus dos opciones, al igual que el segundo y el tercer cromosoma, entonces los posibles resultados se expresan de la siguiente forma:

$(k_1)(k_2)(k_3) = (2)(2)(2) = 2^3 = 8$ posibles resultados

Para diferenciar a cada cromosoma llamaremos al primero como “Largo” y su respectivo par será L1 y L2, al siguiente como “Mediano” y su respectivo par M1 y M2, y al tercero como “Corto” C1 y C2. Entonces los ocho posibles resultados son:

L_1, M_1, C_1	L_1, M_1, C_2	L_1, M_2, C_2	L_1, M_2, C_1
L_2, M_1, C_1	L_2, M_1, C_2	L_2, M_2, C_2	L_2, M_2, C_1

2.8.3 Permutaciones

Una **permutación** es un arreglo específico de objetos en una secuencia específica. Un caballo (C), una vaca (v) y una oveja (O); podrían ser ordenadas linealmente dentro de sus seis formas distintas: CVO; COV; VCO; VOC; OCV; OVC. Esta serie de resultados puede examinarse al observar que existen tres posibles resultados para la primera posición, y solo dos opciones para que un animal ocupe la segunda posición; y por último solo una opción para que solo un animal (el restante) ocupe la tercera posición. Por lo tanto $K_1=3$, $K_2=2$, $K_3=1$, así que los posibles arreglos lineales estarán dados por: $(k_1)(k_2)(k_3)=(3)(2)(1)=6$ maneras de ordenar linealmente los tres animales. También podríamos decir que hay seis permutaciones de tres objetos diferentes.

En general si hay ***n*** posiciones lineales para que sean llenados con ***n*** objetos, la primera posición será ocupada por alguno de los ***n*** objetos, la segunda posición será ocupada por ***n–1*** objetos, la tercera posición por ***n–2*** objetos, y así sucesivamente hasta la última posición, la cual será ocupada por una sola opción-objeto. Así el llenado de ***n*** posiciones con ***n*** objetos resulta en ${}_nP_n$ permutaciones donde:

$$ {}_nP_n = n(n-1)((n-2)(n-3) = (3)(2)(1) = 6 $$

Lo cual puede ser descrito de forma mas simple en notación factorial (!) como sigue:

$$ {}_nP_n = n! $$

Ejemplo 4. ¿En cuantas secuencias pueden seis diapositivas ser mostradas a través de un proyector?

$$ {}_6P_6 = 6! = (6)(5)(4)(3)(2)(1) = 720 \text{ posiciones distintas} $$

Si uno tiene ***n*** objetos, pero con un menor número de posiciones para colocarlos, entonces habría un considerable menor número de formas para colocar los objetos que en el caso donde hay posiciones para todas las ***n***. Por ejemplo hay ${}_4P_4 = 4! = 24$ formas de acomodar linealmente un caballo (C), una vaca (V), una oveja (O) y un puerco (P) en cuatro posiciones, Sin embargo solo hay doce formas de acomodar linealmente dos de esos cuatro animales:

C, V; C, O; C, P; V, C; V, O; V, P; O, C; O, V; O, P; P, C; P, V; P, O

Por lo tanto para calcular el número de **permutaciones lineales** de ***n*** objetos tomados ***x*** a la vez:

$$ {}_nP_x = \frac{n!}{(n-x)!} $$

Para el ejemplo mencionado arriba tenemos:

$$ {}_4P_2 = \frac{4!}{(4-2)!} = \frac{4!}{2!} = \frac{(4)(3)(2)(1)}{(2)(1)} = 12 $$

La ecuación previa ${}_nP_n = n!$ Es un caso particular de la ecuación ${}_nP_x = n!/(n-X)!$ Donde $x = n$; por lo que es importante destacar que $0! = 1$.

Ejemplo 5. En cuantas formas posibles puede una secuencia de cuatro diapositivas formarse de seis diapositivas?

$$ {}_nP_x = {}_4P_2 = \frac{6!}{(6-4)!} = \frac{6!}{2!} = \frac{(6)(5)(4)(3)(2)(1)}{(2)(1)} = 12 $$

Si alguno de los objetos son indistinguibles, por ejemplo en el caso de tener: dos caballos (C), una vaca (V), una oveja (O) el número de permutaciones de los cuatro animales sería de 12:

CCVO; CCOV; CVCO; CVOC; COCV; COVC

VCCO; VCOC; VOCC; OCCV; OCVC; OVCC

Entonces si n_i, representa el número de individuos en la categoría ***i*** (en este caso el número de animales en la especie ***i***); entonces en este ejemplo $n_1 = 2, n_2 = 1, n_3 = 1$, y escribimos el número de permutaciones a través de la siguiente notación:

$$ {}_nP_{n1\,n2\,n3} = \frac{n!}{n_1!\,n_2!\,n_3!} = \frac{n!}{n_1!\,n_2!\,n_3!} = \frac{4!}{2!\,1!\,1!} = 12 $$

Si los cuatro animales fueran dos caballos (C), y dos vacas (V), entonces habría solamente seis permutaciones.

En general, si n_1 miembros de la primera categoría de objetos es indistinguible, como lo es n_2 de la segunda, y como lo es n_3 de la tercera, y así sucesivamente hasta n_k miembros de la ***k-ésima*** categoría, entonces el número total de permutaciones diferentes es:

$$ {}_nP_{n1,\,n2,\ldots nk} = \frac{n!}{n_1!\,n_2!\ldots n_k!} = \frac{n!}{\prod_{i=1}^{k} n_1!} $$

Ejemplo 6. De un total de doce plantas, seis son de una primera especie, cuatro de la segunda, y dos de la tercera especie. ¿Cuántas diferentes secuencias de especies son posibles?

$$ {}_nP_{n1,\,n2,\ldots nk} = {}_{12}P_{6,4,2} = \frac{12!}{6!4!2!} = 13{,}860 $$

A manera de resumen podemos decir que las permutaciones se consideran grupos de objetos o elementos donde la secuencia dentro de los grupos fue lo importante. Sin embargo en otros casos, solo los componentes de un grupo, no su arreglo dentro del grupo es lo importante. Anteriormente, se señaló que si seleccionamos dos animales de un grupo de cuatro animales, existen doce formas de arreglarlos linealmente en grupos de dos animales. Sin embargo algunos de esos arreglos contienen exactamente la misma clase de animales solo que en orden distinto.

2.8.4. Combinaciones

En el caso antes señalado, si los grupos de animales es lo importante para nosotros, pero no la secuencia de objetos dentro de los grupos, entonces estamos hablando de combinaciones, en lugar de permutaciones. La notación del número de combinaciones de n objetos tomados ***X*** a la vez es ${}_nC_x$, y el cálculo es de la siguiente forma:

$$ {}_n C_x = \frac{{}_nP_x}{X!} = \frac{n!}{X!\,(n - X)\,!} $$

Así en el ejemplo en donde se señala que hay solo doce formas de acomodar linealmente dos de estos cuatro animales (un caballo (C), una vaca (V), una oveja (O) y un puerco (P)), solo tenemos seis formas de combinarlos: $n_1 = 4$ y $n_2 = 2$ tenemos la siguiente expresión :

$$ {}_4 C_2 = \frac{4!}{2!\,(4 - 2)\,!} = \frac{4!}{2!\,2!} = \frac{(4)(3)(2)(1)}{(2)(1)(2)(1)} = \frac{(4)(3)}{2} = 6 $$

Es decir, las siguientes seis combinaciones en grupos de dos de la siguiente forma:

CV; CO; CP; VO; VP; OP

Ejemplo 7. De un total de diez perros, ocho son usados en un experimento de laboratorio. ¿Cuántas combinaciones diferentes de ocho animales pueden ser tomadas de un total de diez?

$$ {}_{10} C_8 = \frac{10!}{8!(10 - 8)!} = \frac{10!}{8!2!} = \frac{(10)(9)(8)(7)(6)(5)(4)(3)(2)(1)}{(8)(7)(6)(5)(4)(3)(2)(1)(2)(1)} = \frac{(10)(9)}{2} = 45 $$

Es importante hacer notar que ${}_nC_n = 1$ y ${}_nC_1 = n$; así también es importante señalar que hay ***n*** formas de seleccionar ***n*** elementos a la vez, expresado como ${}_nC_x = {}_nC_{n-x}$.

2.8.5. Conjuntos

Un conjunto está definido como una colección de elementos. Por ejemplo, un conjunto puede ser un grupo de cuatro animales, una colección de ocho aminoácidos, un salón de 25 estudiantes, etc. Si un conjunto incluye cuatro elementos H, C, S, P; y un segundo conjunto consiste de los elementos P, S, H, C; entonces decimos que los conjuntos son iguales debido a que contienen los mismos elementos. Si otro conjunto consiste de los elementos H y P, este se declararía como un subconjunto de H, C, S, P.

Por lo tanto la determinación de combinaciones ***X*** elementos tomados de un total de ***n*** elementos es la contabilización de las posibles subconjuntos de elementos de un conjunto de ***n*** elementos.

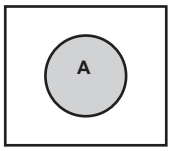
Si en un experimento hay un conjunto de posibles resultados; podemos referirnos a este conjunto como el conjunto **resultante** o **espacio de muestra**. Cada elemento de un conjunto es uno de los posibles resultados del experimento. Por ejemplo si un experimento consiste en lanzar dos monedas el

conjunto resultante esta constituido por cuatro elementos: HH; HT; TH; TT, debido a que estos son todos los posibles resultados.

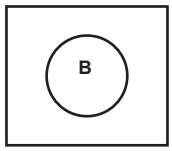
Un subconjunto de un conjunto resultante es denominado un **evento o realización de un fenómeno**. Si el conjunto resultante fuesen por ejemplo los lados de un dado: 1, 2, 3, 4, 5, 6; entonces un evento podría ser “números pares” (2, 4, 6) y otro evento podría ser “números mayores a cuatro” (5, 6). En el caso de tener dos monedas un evento podría ser “Ambas monedas distintas” (T, H; y H, T) y otro evento podría ser que ambas fueran de un solo lado (T, T o H, H).

Si dos eventos en una serie resultante tienen algunos elementos en común, entonces los dos eventos se encuentran en la **intersección** de dos eventos; que es aquel subconjunto compuesto por aquellos elementos comunes entre las dos series resultantes.

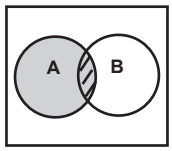
Por ejemplo



A



B

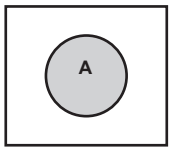


A∩B

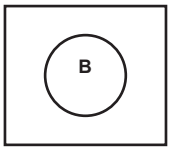
Por ejemplo un evento con los números pares de un dado (2, 4, 6) y el evento con números mayores a cuatro en un dado (5, 6), tienen un elemento en común que es el 6. Por lo tanto el número seis es la intersección de los dos eventos. Así para los números pares de un dado (2, 4, 6) y los números menores a cinco (1, 2, 3, 4), la subserie de intersección consiste en los elementos comunes entre ambos eventos, o sea 2 y 4.

Si dos eventos no tienen elementos en común se dice que estos eventos son mutuamente **excluyentes o exclusivos**, por lo tanto los conjuntos resultantes están desunidos. Por ejemplo, el evento con números pares de un dado con respecto al evento con números impares de un dado son mutuamente exclusivos y no hay elementos comunes para ambas series.

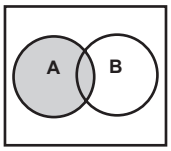
Los elementos de un y otro o de ambos conjuntos resultantes se denominan unión entre los dos eventos o conjuntos de elementos. La unión de dos eventos como los números 1, 2, 3, 4, 6. de un evento y los 5, 6 de otro comparten el elemento 6 en ambos conjuntos, a esto se le llama conjuntos **no ajenos**.



A

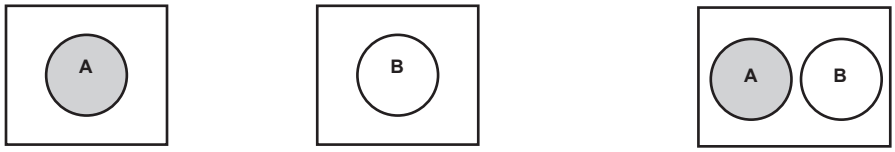


B



AUB

Pero si los eventos o conjuntos contienen elementos diferentes, su unión representará la unión de conjuntos ajenos. Por ejemplo si un evento consiste de la serie de números pares (2, 4, 6) y otro de la serie (1, 3, 5) la unión de estos conjuntos corresponderá al conjunto (1,2,3,4,5,6).



A B AUB

2.8.6. Cálculo de Probabilidad de un Evento

Como se definió previamente la frecuencia relativa de un evento es la proporción del total de observaciones de resultados que un evento representa. Considera una serie resultante con dos elementos, tal como los posibles resultados de lanzar una moneda (H, T) ó el sexo de una persona (♀ o ♂). Si ***n*** es el número total de lanzamientos de una moneda y ***f*** es el número total de "águilas" observadas, entonces la frecuencia relativa de águilas es observadas ***f/n.*** Así si el lado "águila" de una moneda fue observado 52 veces de 100 lanzamientos de una moneda, la frecuencia relativa es 52/100 = 0.52 (52%). Si 275 hombres ocurren en 500 nacimientos humanos, la frecuencia relativa de hombres es ***f/n*** = 275/500 = 0.55 (55%).

La probabilidad de un evento es la verosimilitud (Relación entre dos probabilidades) de ese evento, expresado ya sea como la frecuencia relativa observada de un número grande de datos o por el conocimiento del sistema de estudio.

Ejemplo 8. Una muestra de 852 vertebrados es tomada al azar en un determinado bosque, obteniéndose los siguientes resultados.

Subseries de vertebrados	Número	Frecuencia relativa
Anfibios	53	0.06
Tortugas	41	0.05
Serpientes	204	0.24
Aves	418	0.49
Mamíferos	136	0.16
Total	852	1.00

En este ejemplo las frecuencias relativas del grupo de vertebrados fue calculada de una muestra tomada al azar. Si consideramos el presente ejemplo, nosotros podríamos suponer que cada animal tiene la misma oportunidad de ser capturado. Sin embargo, si nosotros consideramos la frecuencia relativa como un valor de probabilidad, ***P***; para el caso particular del grupo serpiente su probabilidad de ser capturado es de P(Serpientes) = 0.24.

De esta manera, una probabilidad algunas veces puede ser predicha basándonos en el conocimiento del sistema.

2.8.7 Adición de probabilidades

Si dos eventos son mutuamente excluyentes, entonces la probabilidad de que ocurra el evento ***A*** o ***B***, es la suma de las probabilidades de ambos eventos:

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

Por ejemplo si la probabilidad de que al lanzar una moneda el resultado de que caiga sol (o sea un solo de los dos lados) es ½ y de que salga águila (o sea el lado contrario) es también ½, entonces la probabilidad de que el resultado sea cualquiera de los dos lados es:

P (águila o sol)= P(águila) + P(sol) = ½ + ½ = 1.0, es decir que el resultado de águila es 0.5 y de sol es también 0.5, entonces la probabilidad total es 1.

Para el ejemplo de los datos de grupos de vertebrados la probabilidad de capturar un reptil al azar sería:

P(Tortuga o serpiente) = P(Tortuga) + P(serpiente) = 0.05 + 0.24 = 0.29

La regla para la adición de probabilidades para más de dos eventos mutuamente excluyentes, puede ser explicada por ejemplo si consideramos lanzar un dado; en donde la probabilidad de obtener un dos será 1/6, de obtener un cuatro será 1/6, y de obtener un seis será también 1/6. Así la probabilidad de obtener cualquiera de estos tres resultados será:

P(2, 4, 6) = P(2) + P(4) + P(6) = 1/6 + 1/6 +1/6 = 1/2

Es decir que la probabilidad de tener dos será 0.16666667 y lo mismo para obtener 4 y seis, entonces:

P(2, 4, 6) = P(2) + P(4) + P(6) = 0.16666667 + 0.16666667 + 0.16666667 = 0.5

Por otra parte si dos eventos no son mutuamente excluyentes, es decir, existe una intersección entre ambos eventos, entonces la adición de probabilidades deberá ser modificada. Por ejemplo si se lanza un dado la probabilidad de obtener números impares es:

P(números pares) =P(1 ó 3 ó 5) = P(1) + P(2) + P(3) = 1/6 + 1/6 + 1/6 = ½

Y la probabilidad de obtener un número menor a 4 es:

P(número <4) = (1 ó 2 ó 3) = P(1) + P(2) + P(3) = 1/6 + 1/6 +1/6 = ½

Debido a que los dos elementos (1 y 3) se presentan en ambos eventos, el subconjunto 1, 3 es la intersección entre ambos eventos, entonces la estimación de la probabilidad es de la manera siguiente:

P(A o B) = P(A) + P(B) – P(∩AB)

Entonces:

P (Número impar o número <4)= P(Número impar) + P(número <4) – P(∩Número impar o número <4)

P([1, 3, 5] o [1, 2, 3]) – P(1, 3) = [P(1) + P(3) + P(5) + P(1) + P(2) + P(3)] – [P(1) + P(3)] = (1/6 + 1/6 + 1/6) + (1/6 + 1/6 + 1/6) – (1/6 + 1/6) = 4/6 = 2/3

En el caso de tener tres eventos que no son mutuamente excluyentes la situación es más compleja. Como se observa en la figura existen tres interacciones de dos vías (mostradas con líneas verticales) las cuales son A y B; A y C; B y C, así como una sola interacción de tres vías señalada con líneas horizontales (A, B, C) a este tipo de representaciones se les llama diagramas de Venn.

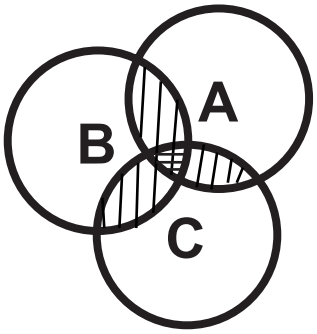


Diagrama de Venn

Si se adicionan las probabilidades de los tres eventos ***A, B*** y ***C*** como ***P(A) + P(B) + P(C)***; se están adicionando las interacciones de dos vías doblemente. Así que se debe sustraer ***P(A∩B), P(A∩C)***, y ***P(B∩C)***. Como también la interacción de tres vías es adicionada tres veces de la adición ***P(A) + P(B) + P(C)*** y sustraídas tres veces a restar las interacciones de dos vías; por la tanto ***P(A∩B∩C)***, debe ser adicionada al final. Por lo tanto para tres eventos que no son mutuamente excluyentes tenemos:

$$P(A, B, C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$$

2.8.8. Multiplicación de Probabilidades

Si dos o más eventos se intersectan, la probabilidad asociada con la intersección es el producto de las probabilidades de los eventos individuales, la cual puede ser denotada de la siguiente forma:

$$P(\cap AB) = [P(A)]\,[P(B)]$$

$$P(\cap ABC) = [P(A)]\,[P(B)]\,[P(C)]$$

Ejemplo 9. la probabilidad de que al lanzar una moneda se obtenga “águila” es ½, sin embargo, si dos monedas son lanzadas la probabilidad de obtener dos “águilas” es:

$$P(\text{Dos águilas}) = [P(\text{Agulia})]\,[P(\text{Aguila})] = (1/2)(1/2) = 0.25$$

Esto puede ser verificado mediante la examinación de la serie resultante:

H,H; H,T; T,H; T,T

Donde la P(H, H) es un resultado de las cuatro igualmente posibles resultados. Así ahora lanzamos tres monedas en vez de dos, la probabilidad de obtener dos águilas será:

$$P(H, H, H) = [P(H)][P(H)][P(H)] = (1/2)(1/2)(1/2) = 0.125$$

2.8.9. Probabilidad Condicional

En muchas ocasiones, la probabilidad de que ocurra un evento, depende de lo que ha ocurrido con otro evento. En este caso se tiene la llamada probabilidad condicional. La probabilidad condicional de ***A***, dado que ha ocurrido el evento ***B***, se escribe ***P(A/B)***; o sea es la probabilidad de que ocurra un evento ***A*** cundo se conoce cierta información relacionada con la ocurrencia de otro evento ***B***.

P(A/B) Probabilidad de que ocurra ***A***, dado que ***B*** ha ocurrido

P(B/A) Probabilidad de que ocurra ***B***, dado que ***A*** ha ocurrido

$$P(A/B) = \frac{P(A \cap B)}{P(B)}$$ *es la probabilidad condicional de A*

$$P(B/A) = \frac{P(B \cap A)}{P(A)}$$ *es la probabilidad condicional de B*

P(A∩B), es la probabilidad conjunta porque denota la intersección de los eventos ***A*** y ***B***.

P(B∩A), es la probabilidad conjunta porque denota la intersección de los eventos ***B*** y ***A***.

P(A) y ***P(B)***, se denominan probabilidades marginales.

Ejemplo 10. La tabla a continuación, presenta el ascenso a catedráticos de los profesores de una institución durante 5 años.

T**abla de acenso al rango de catedrático**

	Hombres (H)	Mujeres (M)	Total
Ascendido (A)	278	26	304
No ascendido (A´)	662	194	856
Total	940	220	1160

Tabla de probabilidad conjunta

	Hombres (H)	Mujeres (M)	Total
Ascendido (A)	0.24	0.02	0.26
No ascendido (A´)	0.57	0.17	0.74
Total	0.81	0.19	1.0

A partir de esta información, se construye una tabla de probabilidades conjuntas, cuyos valores de probabilidad son calculados y aparecen al interior de la tabla.

A este tipo de tablas se les llama "**tablas de probabilidad conjunta o condicional**" y a los totales se les llama "**probabilidad marginal**".

Cual es la probabilidad de que un profesor seleccionado al azar sea hombre (H) y fuera ascendido?

P(H∩A) = 278/1160 = 0.24

Cual es la probabilidad de que un profesor seleccionado al azar sea hombre (H) y no fuera ascendido?

P(H∩A') = 662/1160 = 0.57

Cual es la probabilidad de que un profesor seleccionado al azar sea mujer (M) y fuera ascendido?

P(M∩A) = 26/1160 = 0.02

►

Continuación ejemplo 10

Cual es la probabilidad de que un profesor seleccionado al azar sea mujer (***M***) y no fuera ascendido?

P(H∩A') = 194/1160 = 0.17

Ahora se pueden calcular las probabilidades conjuntas

a) Probabilidad de que un profesor elegido al azar sea ascendido dado que es hombre (H)

P(A/H) = 278/940 = 0.30

Alternativamente

P(A∩H) = P (H)*P(A/H)

Donde:

$$P(A/H) = \frac{P(A \cap H)}{P(H)} = 0.24/0.81 = 0.30$$

b) Probabilidad de que un profesor elegido al azar sea ascendido dado que es mujer (***M***)

P(A/M) = 26/220 = 0.12

Alternativamente

P(A∩M)= P (M)*P(A/M)

$$P(A/M) = \frac{P(A \cap M)}{P(M)} = 0.02/0.19 = 0.12$$

3 Distribución de Probabilidades

3.1. Introducción

En estadística la **distribución de probabilidad** ***F (x)*** es una función de la probabilidad que representa los resultados que se van obteniendo en un experimento aleatorio (figura 10).

Para un número dado ***x***, la probabilidad ***P(X≤ x)*** es:

$$F(x) = P(X \leq x)$$

A ***F(x)*** se le denomina función de distribución de probabilidad de la variable ***X*** y representa la probabilidad de que la variable tome el valor desde - ∞ hasta ***x***.

Literatura sugerida:

http://es.wikipedia.org/wiki/Funci%C3%B3n_de_densidad

Scherrer B., 1984, Biostatistique. Gaëtan Morin Editor. Montreal, Paris, Casa Blanca. 850 p (Pág 221-244).

Zar. J. H., 1999. Bostatistical Analysis (4 edición). Prentice Hall. Estados Unidos. 663 p. (Pag. 521).

http://personal5.iddeo.es/ztt/Tem/t21_distribucion_normal.htm

http://www.bioestadistica.uma.es/libro/node38.htm

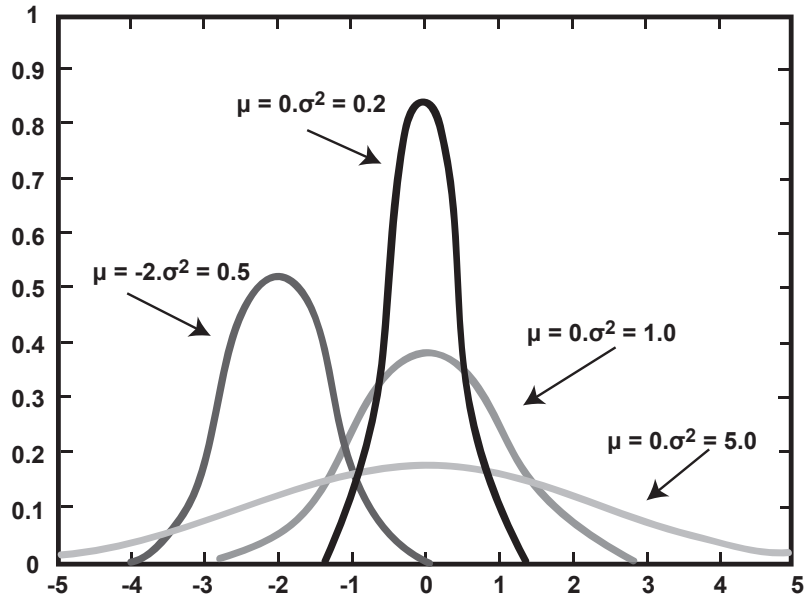


Figura 10. Función de densidad para la distribución normal, modificado de: http://es.wikepedia.org/wiki/Funci%C3%B3n_de_densidad

También se puede definir como la acumulada de la función de densidad de probabilidad, esta última más comúnmente conocida como función de densidad

La **función de densidad** se utiliza en estadística con el propósito de conocer como se distribuyen las probabilidades de un evento en relación al resultado del evento. En este caso se llama función de densidad de probabilidad

Matemáticamente la ***FDP*** (función densidad de probabilidad) es la derivada de la función de distribución de probabilidad.

Las propiedades de ***FDP*** (a veces visto como PDF del inglés) son:

- ***FDP(x)*** ≤ 0.
- La integral de ***FDP(x)*** en el rango especificado para la función, siempre es 1.

Algunas ***FDP*** están declaradas en rangos de - ∞ a ∞, como la de la distribución normal

Para dos números reales cualesquiera ***a*** y ***b*** tal que (***a < b***), los sucesos (***X ≤ a***) y (***a <X ≤ b***) serán mutuamente excluyentes y su suma es el suceso (***X ≤ b***) por lo que tenemos entonces que:

$$P(X \leq b) = P(X \leq a) + P(a < X \leq b)$$

$$P(a < X \leq b) = P(X \leq b) - P(X \leq a)$$

y finalmente

$$P(a < X \leq b) = F(b) - F(a)$$

Por lo tanto una vez conocida la función de distribución ***F(x)*** para todos los valores de la variable aleatoria ***x*** conoceremos completamente la distribución de probabilidad de la variable.

Como la probabilidad es siempre un número positivo, la función de distribución será una función no decreciente que cumple lo siguiente:

$$\lim_{n \to \infty} F(x) = 1$$

Es decir la probabilidad de todo el espacio muestral es 1 tal y como establece la teoría de la probabilidad y por otra parte:

$$\lim_{n \to \infty} F(x) = 0$$

Es decir la probabilidad del suceso nulo es cero.

Para realizar cálculos es más cómodo conocer las distribución de probabilidad, para ver una representación gráfica de la probabilidad es más práctico el uso de la función de densidad.

Existen dos tipos de distribución de variables, las discretas y las continuas.

Se denomina **variable discreta** a aquella que sólo puede tomar unos determinados valores, el conjunto de valores que toma ***X*** es finito o numerable. En este caso la distribución de probabilidad es la suma de la función de densidad, por lo que tenemos entonces que:

$$F(x) = P(X \leq x_i) = \sum_{k=1}^{i} f(x_i)$$

Tal como corresponde a la definición de distribución de probabilidad, esta expresión representa la suma de todas las probabilidades desde — ∞ hasta el valor x_i.

Es importante señalar que muchos autores señalan que en las variables discretas no se denomina "función de densidad" sino "función de probabilidad", y se nota $P_x(x)$ en vez de $f_x(x)$. Dentro de las distribuciones discretas se encuentran principalmente la distribución binomial y de poisson.

Se denomina **variable continua** a aquella que puede tomar cualquiera de los infinitos valores existentes dentro de un intervalo finito. En el caso de variable continua la distribución de probabilidad es la integral de la función de densidad, por lo que tenemos entonces que:

$$F(x) = P(X \leq x_i) = \int_{-\infty}^{x_i} f(x_i)\, dx$$

Dentro de las distribuciones continúas, se tratarán: la distribución normal, *t*-student, chi-cuadrado y Fisher-Snedecor (*F*).

3.1.1 Ley Binomial o Distribución Binomial

Es una de las distribuciones frecuentemente encontrada en estadística aplicada (Jacob Bernoulli, 1713); a partir de esta se pueden analizar datos cualitativos provenientes de una población compuesta de 2 categorías de elementos (variables de conteo, proporciones y porcentajes). **La ley binomial se define como una distribución discontinua que da las probabilidades de ver aparecer un evento de probabilidad p respectivamente 0, 1, 2, 3,.....*i*., n veces a lo largo de n pruebas (sorteos o experiencias) idénticas e independientes. Con esta ley dos eventos pueden aparecer en cada prueba**.

1) probabilidad de éxito (o que ocurra) ***p***

2) probabilidad de fracaso (o que no ocurra) ***q=1-p***, entonces ***p+q =1***

Esto corresponde por ejemplo a la aparición de individuos machos o hembras, blancos o negros, muertos o vivos que presenten una u otra característica. La **distribución binomial se designa con la letra ß(n,p) donde n representa los eventos y p la probabilidad de uno de los eventos**. Así por ejemplo en una familia de ***n*** hijos, cual es la probabilidad de tener ***x*** varones? En este ejemplo, el efecto del nacimiento de un niño corresponde a un evento en donde el resultado corresponde a los dos eventos siguientes: o el hijo es un niño o es una niña. En cada evento, la probabilidad ***p*** de que nazca un niño es igual a 0.5, y la probabilidad ***q*** de que sea una niña es también 0.5; si la probabilidad de tener un niño no varía de un nacimiento a otro y si hay independencia entre cada evento, la probabilidad de tener ***x*** niños en una familia de ***n*** hijos esta dada por una distribución binomial.

Por otra parte si se quiere observar el comportamiento de ***n*** ratas que penetran a un laberinto en forma de H sucesivamente. ¿Cual es la probabilidad de que ***n*** ratas se dirijan a la parte derecha superior del

laberinto? A cada prueba dos eventos son posibles, o penetra y toma el camino adecuado o no lo toma. Entonces como hay 4 itinerarios posibles, la probabilidad del primer evento es ***p*** = ¼ (0.25) y la del segundo evento es ***q*** = ¾ (0.75). Si las ratas no han sido condicionadas y si la rama inferior derecha no contiene ningún elemento atractivo o repulsivo para afectar ***p*** o ***q***, las probabilidades de ***x*** estarán dadas por una distribución binomial.

En otro ejemplo: si se toma una población de hembras y machos en las proporciones de ***p = 0.4*** y ***q = 0.6*** respectivamente y se toma al azar una muestra de dos individuos de la población. La probabilidad de que la primera sea hembra (0.4) es ***p*** y la probabilidad de que la segunda sea hembra es también ***p.*** la probabilidad de tener dos hembras en una muestra de dos es ***(p)(p) = p^2=(0.4)(0.4) = 0.16***. Por lo tanto la probabilidad de que una muestra consista de dos machos será ***(q)(q) = q^2 =(0.6)(0.6) = 0.36***.

Entonces ¿Cuál sería la probabilidad de que en una muestra ***n*** = 2, se obtenga un macho y una hembra respectivamente? Esto podría ocurrir con el primer elemento siendo macho y el segundo hembra (probabilidad ***pq***) o con el primer elemento siendo hembra y el segundo macho (probabilidad ***qp***). La probabilidad de cualquiera de los resultados mutuamente excluyentes, es la suma de las probabilidades de cada resultado, así la probabilidad de tener en la muestra una macho y una hembra es ***pq + qp*** = 2 ***pq*** = (2)(0.4)(0.6) = 0.48. En este contexto es importante destacar que 0.16 + 0.36 + 0.48 = 1.00, lo cual nos indica que nosotros solo podemos obtener tres posibles resultados asociados a tres probabilidades con ***n*** = 2 donde ***p*** = 0.4, y ***q*** = 1 **- *p***, donde las probabilidades en suma resultan en el 100% de la probabilidad. Así por ejemplo para ***n*** = 5, las distribuciones de probabilidad se muestran en la figura 11.

Así también, las estimaciones de probabilidades binomiales pueden también ser calculadas para eventos donde ***p***= 0.5, ***q***= 0.5 ***n***= 5, ***p***=0.3 y ***p***= 0.7 o bien ***p***= 0.1 y ***q***= 0.9 aquí observamos que a diferencia de cuando teníamos ***p***= 0.5 y ***q***= 0.5 con una distribución simétrica de probabilidades, al cambiar la fre-

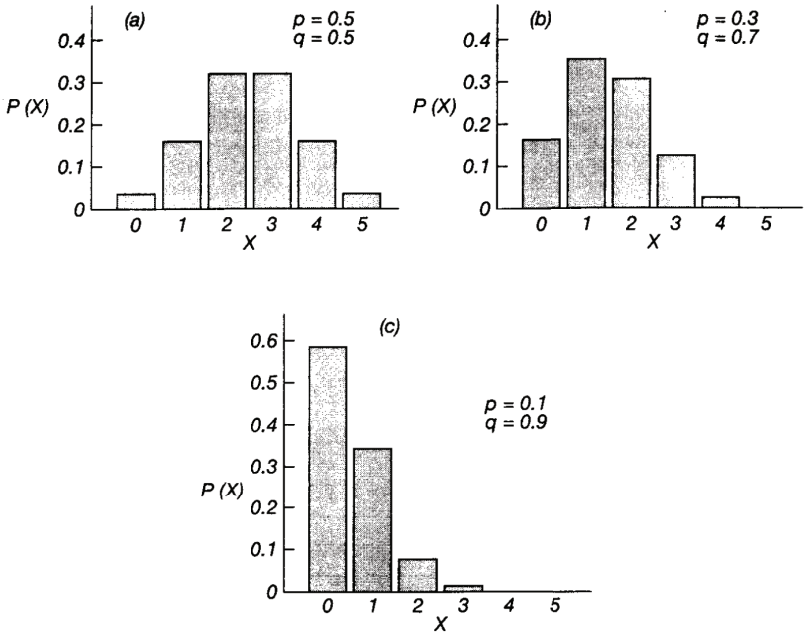


Figura 11. Distribuciones de probabilidad binomial con ***n*** = 5
(Tomada de Zar (1999) pág. 521)

cuencia binomial de forma desigual la probabilidad se distribuye de manera asimétrica o sesgada como se observa en la figura 11b. Se puede hablar ahora de que las probabilidades de aparición están dadas por la ley de Newton (***p+q***).

x	P(x) Probabilidad de x
0	$C_n^0 q^n$
1	$C_n^1 q^{n-1} p$
2	$C_n^2 q^{n-2} p^2$
3	$C_n^3 q^{n-3} p^3$
i	$C_n^i q^{n-i} p^i$
n-1	$C_n^{n-1} q p^{n-1}$
n	$C_n^n p^n$

Términos sucesivos del desarrollo del binomio de Newton-Distribución de Probabilidad: ***β(n,p)***.

Los coeficientes designan el número total de combinaciones de n objetos tomados de ***x*** en ***x***. Para determinarla solamente hay que calcular el número total de combinaciones según la siguiente fórmula:

$$C_n^x = \frac{n!}{x!(n-x)\ !}$$

Así si ***n***= 1, el binomio de Newton se escribe ***(p+q)¹*** y su desarrollo es igual a ***p+q***, su tabla de probabilidad es: ***β(1,p)***

X	p(x)
0	q
1	p

Si ***n***= 2, la ley binomial se escribe ***(p+q)²*** y su desarrollo es igual a $p^2+2pq+q^2$, y su tabla de probabilidades sería: ***β(2,p)***

X	p(x)
0	q^2
1	2pq
2	p^2

Finalmente si **n= 7**, la distribución de probabilidad será: β***(7,p)***

Por tanto la probabilidad de ver aparecer ***x*** veces un evento de probabilidad ***p*** en ***n*** eventos idénticos e independientes se escribe:

$$p(x) = C_n^x q^{n-x} p^x = \frac{n!}{(n-x)!\ x!}\ q^{n-x} p^x$$

X	p(x)
0	$\frac{7!}{0!\ (7-0)!}\ q^7 = q^7$
1	$\frac{7!}{1!\ (7-1)}\ q^6 p = 7q^7 p$
2	$\frac{7!}{2!\ (7-2)}\ q^5 p^2 = 21q^5 p^2$
3	$\frac{7!}{3!\ (7-3)}\ q^4 p^3 = 35q^4 p^3$
4	$\frac{7!}{4!\ (7-4)}\ q^3 p^4 = 35q^3 p^4$
5	$\frac{7!}{5!\ (7-5)}\ q^2 p^5 = 21q^2 p^5$
6	$\frac{7!}{6!\ (7-6)}\ q p^6 = 7q p^6$
7	$\frac{7!}{7!\ (7-7)}\ p^7 = p^7$

Ejemplo 11. En una familia de ***n*** hijos, cual es la probabilidad de tener ***x*** niños? Si el número de hijos es ***n***=1 y si la probabilidad de tener 1 niño en 1 nacimiento es ***p***= 0.5. La probabilidad de tener un niño en una familia de 1 hijo es: ***β***(1,0.5)

$$P(1) = \frac{1}{1!\ 0!} * 0.5^0 * 0.5^1 = 0.5$$

Recordar que 0! = 1 y que cualquier número elevado a la 0 potencia es 1

Y la probabilidad de una niña es:

$$P(1) = \frac{1}{1!\ 0!} * 0.5^0 * 0.5^1 = 0.5$$

Así en una familia de dos hijos, la distribución de probabilidad es: β(2, 0.5)

X	p(x)
0	$\frac{2!}{0!(2-0)!}\ 0.5^2 * o.5^0 = 0.25$
1	$\frac{2!}{1!(2-1)!}\ 0.5 * o.5 = 0.5$
2	$\frac{2!}{2!(2-2)!}\ 0.5^0 * o.5^2 = 0.25$

Una vez obtenidas las probabilidades, estas pueden ser representadas en una gráfica de probabilidad (Fig. 12). Es necesario saber que actualmente existen tablas de probabilidad para diferentes valores de ***n, x*** y ***p***, lo que hace innecesario el cálculo de las mismas (Tabla 1 anexa).

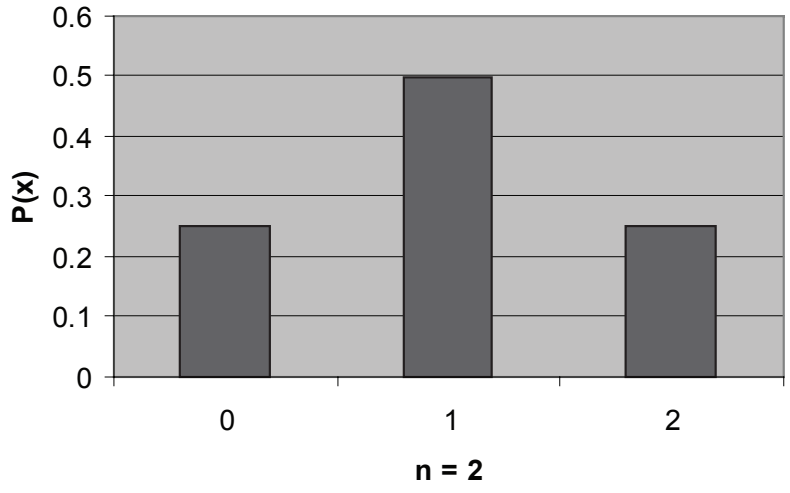


Figura 12. Distribución de probabilidad de una familia de 2 hijos β(2, 0.5).

3.1.1.1 Los momentos de la distribución binomial

La esperanza matemática. La esperanza de una distribución de probabildad es:

$$E(X) = \sum_{i=1}^{n} P(Xi * Xi$$

Para la ley binomial esto corresponde a : ***E(x) = np***, así la esperanza matemática de **x** número de varones en una familia de 7 hijos es:

E(X) = 7*0.5 = 3.5 niños (0.5 es la probabilidad de niños en cada evento).

Varianza. La varianza de una distribución de probabilidad se escribe:

$$\sigma^2 = E(x^2) - \mu^2$$

Si la distribución obedece a una ley binomial esta es igual a:

$$\sigma^2 = npq \text{ y la desviación estándar } \sigma = \sqrt{\sigma^2}$$

Así la varianza del número de varones en una familia de 7 hijos es:

$$\sigma^2 = 7*0.5*0.5 = 1.75 \text{ y la desviación estándar } \sigma = \sqrt{1.75} = 1.32$$

El coeficiente de asimetría (α_3). Este se calcula a partir del momento de tercer orden, en una distribución de probabilidad α_3 se escribe:

$$\alpha_3 = \frac{E(X - \mu)^3}{(\sqrt{E(X - \mu)^2})^3}$$

Si la distribución de probabilidad obedece a una distribución binomial, el coeficiente de asimetría será:

$$\alpha_3 = \frac{q - p}{\sqrt{npq}}$$

Como el denominador ($\sqrt{\boldsymbol{npq}}$) siempre es positivo, el signo de α_3 depende de ***p*** y ***q***, entonces tres posibilidades pueden aparecer:

Si $\boldsymbol{p > q} \Rightarrow \alpha_3 < 0$ asimetría izquierda

Si $\boldsymbol{p = q} \Rightarrow \alpha_3 = 0$ curva simétrica

Si $\boldsymbol{p < q} \Rightarrow \alpha_3 > 0$ asimetría derecha

Así por ejemplo de la familia de 7 hijos, $\alpha_3 = \frac{0.5 - 0.5}{\sqrt{7 * 0.5 * 0.5}} = 0.189$

lo que indica que esta distribución de probabilidad presenta una asimetría a la derecha.

El coeficiente de aplanamiento o curtosis (α_4). Este momento de cuarto orden se calcula para una distribución normal como:

$$\alpha_4 = \frac{E(X - \mu)^4}{(\sqrt{E(X - \mu)^2})^4}$$

Si la distribución de probabilidad obedece a una distribución binomial, el coeficiente de aplanamiento es igual a:

$$\alpha_4 = 3 + \frac{1 - 6pq}{npq}$$

Tomando en cuenta el ejemplo anterior: $\alpha_4 = 3 + \frac{1 - (6*0.5*0.5)}{(7*0.5*0.5} = 2.71$

3.1.2. Ley de Poisson o Distribución de Poisson

La distribución de Poisson es una distribución teórica discontinúa que se deriva de la ley binomial donde uno de los eventos tiene una probabilidad muy baja. Fue descrita en 1830 por Simeon Denis Poisson (1781-1840) un matemático y físico francés (Féron, 1978), aunque Abraham de Moivre (1667-1754) aparentemente la describió previamente en 1718 (David, 1962:1968). Esta distribución se aplica a fenómenos accidentales (de aquí que la probabilidad sea baja ***p***<0.05). Se utiliza cuando se encuentran eventos o individuos distribuidos al azar en el espacio y en el tiempo. Así cuando se cuentan organismos en cuadrantes o parcelas ó en volúmenes se utiliza la ley de Poisson (Pruebas de pureza o germinación de semillas, cuenta de insectos, de hierbas, colonias de bacterias, distribución espacio-temporal de plagas).

En la distribución de Poisson ***n*** siempre es elevado puesto que para ver aparecer un evento raro, hay que hacer muchas pruebas.

Los términos de la distribución de Poisson son:

$$P(x) = \frac{\mu^x}{X!} e^{-\mu}$$

Donde ***P(x)*** es la probabilidad de ***X*** ocurrencias en una unidad de espacio o tiempo y ***μ***, es la media de la población de ocurrencias en una unidad de espacio o tiempo. Así,

$$P(0) = e^{-\mu}$$

$$P(1) = e^{-\mu}\mu$$

$$P(2) = \frac{e^{-\mu}\mu^2}{2}$$

$$P(3) = \frac{e^{-\mu}\mu^3}{(3)(2)}$$

$$P(4) = \frac{e^{-\mu}\mu^4}{(4)(3)(2)}$$

etc., donde ***P(0)*** es la probabilidad de que un evento no ocurra en una unidad de espacio o tiempo, ***P(1)*** es la probabilidad de exactamente una ocurrencia en una unidad de espacio o tiempo, y así sucesivamente. En la figura 13 se muestran diferentes distribuciones de probabilidad de "Poisson" para diferentes valores promedio.

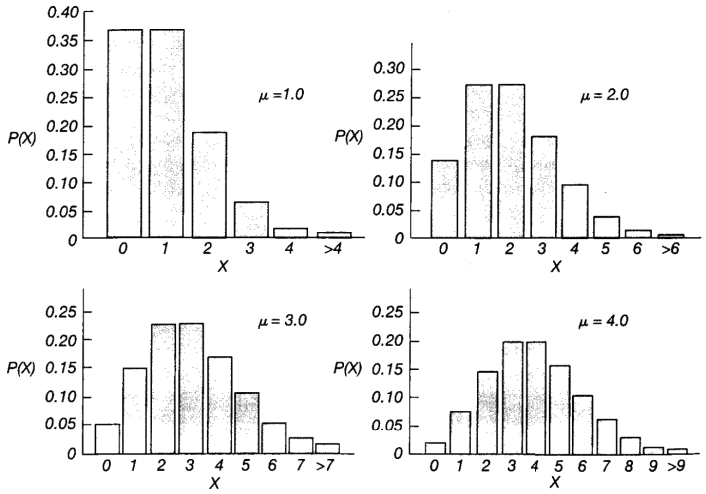


Figura. 13. Distribuciones de Poisson para diferentes valores promedio (tomada de Zar (1999) pág. 572).

Ejemplo 12.Un administrador de un hospital que ha estudiado las admisiones diarias durante años, ha llegado a la conclusión que estas se distribuyen de acuerdo a una ley de Poisson. Las admisiones de emergencia han sido en promedio 3 por día. Encontrar la probabilidad de que de 1 día a otro, ocurran exactamente 2 admisiones de emergencia.

$$P(x) = \frac{\mu^x}{X!}e^{-\mu} = P(x) = \frac{\mu^x e^{-\mu}}{X!} = P(x) = \frac{3^2 e^{-3}}{2!} = \frac{(9) - (0.04978)}{2} = \frac{0.448}{2} = 0.2240$$

y si no ocurriera ninguna admisión: $P(x) = \frac{3^0 e^{-3}}{0!} = 0.049$

Es importante señalar que la tabla de probabilidades ***p(x)*** para ***μ*** se presenta en el anexo, tabla 2.

3.1.2.1 Los momentos de la distribución de Poisson

La esperanza matemática. Esta ley depende de un solo parámetro, la media y utiliza para su cálculo, tablas que dan directamente la distribución de probabilidad de ***X*** para diferentes valores de:

μ : E(X) = np*, así* ***μ*** *=* ***np****, esto es igual que la ley binomial.*

La varianza. Como para la ley binomial, la varianza es $\sigma^2=npq$, sin embargo como ***p*** es muy próximo a cero (recordar que el valor de ***p*** es siempre muy pequeño en una distribución de poisson), la ecuación se simplifica como: $\sigma^2=np(1-p)$

Porque $q= (1-p)$

$\sigma^2=np(1-0)$

$\sigma^2=np$ *y como.* $\mu = np$ *entonces* $\sigma^2= \mu$.
La igualdad de la media y la varianza es una propiedad importante de la ley de Poisson.

La desviación estándar es: $\sigma= \sqrt{\sigma^2}$

El coeficiente de asimetría(α_3)

$$\alpha_3 = \frac{1}{\sqrt{\mu}}$$

Como $\sqrt{\mu}$ no puede ser negativo, la distribución de probabilidad es siempre simética a la derecha.

Coeficiente de aplanamiento o curtosis (α_4)

$$\alpha_4 = 3 + \frac{1}{\sqrt{\mu}}$$

como ***μ*** es siempre positivo, la distribución es menos aplanada que la distribución normal.

Relación entre la ley binomial y la de Poisson. La distribución binomial, tiende hacia la de Poisson cuando ***p*** disminuye y ***n*** aumenta. En la práctica se admite que un evento es raro si su probabilidad es inferior a 0.05. La aproximación de la ley binomial a la de Poisson es satisfactoria si ***n*** es al menos igual a 50.

Ejemplo 13. El ave *Turdus torquatus*, es un pájaro que en otoño se acerca a los bosques de arbustos de montaña entre 1500 y 2000 m de altitud. En 1968, en la estación ornitológica situada en los Alpes Franceses a 1700 m de altitud, 48 aves fueron capturadas con redes "japonesas" durante 89 días en la época de arribo. La distribución de frecuencias se presenta en la siguiente tabla:

Esto implica que ningún ave se capturo durante 56 días, veintidós aves de capturaron por cada día durante 22 días, dos días se capturaron 9 aves cada día, haciendo 18 aves en total, i ave se capturo por día en tres días y cinco aves otro día, por lo que hay 48 aves en total.

¿Sabiendo que el número medio de aves capturadas por día es 0.539, cual es la distribución de probabilidad del número ***x*** de aves capturadas por día?.

Ejemplo 13 (continuación)

Como cada prueba consiste en la captura de un ave, esta captura se produce o no un día ***J*** dado. La probabilidad de que esta se produzca el día ***J***, es igual a 1/89 (***p*** = 1/89 = 0.011), por lo tanto muy baja. El número de capturas es alto puesto que son 48 aves (***n*** = 48). Si las redes no son desplazadas de lugar durante el año, si las aves no se acostumbran a las redes y no las evitan, si su comportamiento no es ni gregario ni territorial, los 48 eventos son idénticos e independientes y la probabilidad de captura se mantiene constante de un evento a otro. En estas condiciones la distribución de probabilidad de ***x*** se calcula como se observa en la tabla para una media ***μ = np*** = 48*(1/89) = 0.539

Número de aves capturadas/día X_i	Número de días en que se capturo f_i
0	56
1	22
2	9
3	1
4	0
5	1
6	0
	Total=89

Ahora la varianza y la desviación estándar se calculan como sigue:

$\sigma^2 = npq$ (como proviene de la binomial)

σ^2=48*1/89*0.988= 0.533 porque ***q = (1-p)***

y entonces (1-0.1123) = 0.988

"observar que la ***μ*** y la σ^2 son similares.

La desviación estándar es igual:

$\sigma = \sqrt{\sigma^2} = \sqrt{0.533} = 0.738$ aves/día

El coeficiente de asimetría:

$\alpha^3 = \frac{1}{\sqrt{0.539}} = 1.362$,

asimetría a la derecha acentuada

y el coeficiente de aplanamiento:

$\alpha 4 = 3 + \frac{1}{0.539} = 4.855$,

la distribución presenta una mayor intensidad que la distribución normal (figura 14).

Tabla de distribución de probabilidad	
X	**P(x)**
0	$e^{-\mu} = 0.583$
1	$P(1) = e^{-\mu} = (0.583)(0.539) = 0.3144$
2	$P(2) = \frac{e^{-\mu}\mu^2}{2} = (0.3144)(0.539)^2/2! = 0.0847$
3	$P(3) = \frac{e^{-\mu}\mu^3}{(3)(2)} = (0.0847)(0.539)^3/3! = 0.0152$
4	$P(4) = \frac{e^{-\mu}\mu^4}{(49)(3)(2)} = (0.0152)(0.539)^4/4! = 0.0021$
5	$P(5) = \frac{e^{-\mu}\mu^5}{(49)(3)(2)} = (0.0021)(0.539)^5/5! = 0.0002$

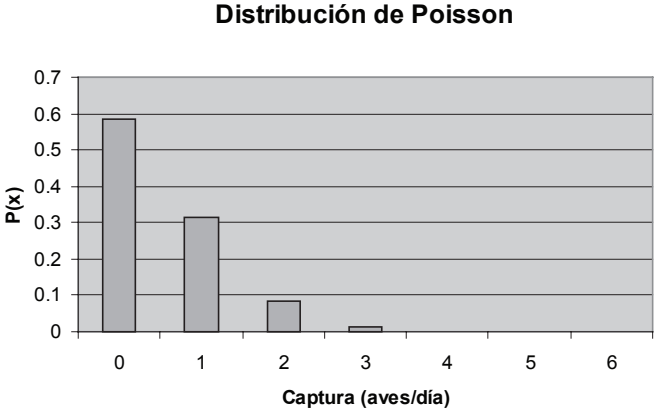


Figura 14. Distribución de Poisson de 48 aves capturadas durante 89 días.

3.1.3. La Distribución Normal, N(μ,σ)

La distribución normal es una distribución continua, juega un papel muy importante tanto en la teoría como en la práctica. Históricamente su importancia fue presentada desde 1733 por Abraham de Moivre, quien trabajaba sobre la ley binomial cuando ***n*** tiende al infinito y ***p*** no llega a 0. En 1772, Laplace la estudia también en su teoría de los errores, sin embargo esta ley adquiere su forma definitiva con Gauss en 1809 y Laplace en 1812, así esta ley se conoce como ley de Laplace, Ley de Gauss o Ley de Laplace-Gauss y como ley normal.

Comúnmente la distribución de frecuencias de datos en escala de intervalo o razón se ha observado que tiene una preponderancia de valores alrededor de la media con progresivamente un menor número de observaciones hacia los extremos del intervalo de valores. Si ***n*** es grande, los polígonos de frecuencias de algunos datos biológicos son de "forma de campana" siendo posible de ser explicados mediante la función de distribución normal o de Gauss.

La distribución normal proviene de la ley binomial:

$$p(x) = \frac{n!}{(n-x)!x!} = q^{n-x}\, p^{x}$$

*con **x** variando entre 0 a **n***

En este momento es conveniente mencionar el **"teorema del límite central"**. Sea X_i la continuación de una variable independiente que presenta la misma distribución, si $\mu = E(X_i)$ y $\sigma^2 = var(x_i)$ y si el muestreo se realiza a partir de una población no distribuida normalmente, este teorema indica lo siguiente "Dada una población de cualquier forma funcional con una media μ y varianza finita σ^2, la distribución muestral de $\overline{X}$, calculada a partir de muestras de tamaño ***n*** de esta población, estará distribuida aproximadamente normal con una media μ y varianza σ^2/n cuando el tamaño de la muestra es grande". Así si en la distribución binomial ***n*** tiende hacia el infinito y ***p*** no esta muy próximo de 0 y 1 (si no tendería a la distribución de Poisson), la distribución tiende hacia la distribución normal. Esto sugiere el hecho de que se puedan tomar muestras a partir de una distribución que no es normal con la garantía de tener buenos resultados, siempre que se tome una muestra grande. Mayores conocimientos sobre los cálculos matemáticos de este teorema pueden ser consultados en las obras de Lebart *et al*. (1979).

Formalmente la función de densidad normal se escribe como:

$$Z = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-(Xi-\mu)^2/2\sigma^2}$$

Donde ***Z*** representa la altura de la ordenada de la curva que representa la densidad de los datos y que es función de la variable ***x***. La ecuación contiene la constantes π (3.14159) y ***e*** (2.7182), así como los parámetros estadísticos σ (Desviación estándar) y μ (Media aritmética). La media paramétrica y la desviación típica paramétrica, determinan la forma y localización de la distribución. Así, para alguna desviación estándar determinada ***(σ)***, existe un número infinito de curvas normales posibles dependiendo del valor de μ. Una curva normal con media igual a cero y desviación estándar igual a uno, se dice que es una curva normal estandarizada o centrada y reducida.

3.1.3.1. Distribución normal centrada y reducida

De acuerdo a la teoría del límite central, la distribución de probabilidad de una distribución normal transformada de la binomial se expresa como sigue:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Esta distribución depende de dos parámetros, la media que localiza a la distribución y la desviación estándar que mide la dispersión de valores en torno a la media. Si se trabaja con una distribución de media 0 y desviación estándar de 1, a este cambio de variables de le llama distribución central normal, centrada y reducida. Para que la media pueda ser 0 el primer cambio de la variable consiste a remplazar ***x*** por ***X*** de la siguiente manera: ***X = x-μ*** esta transformación conduce a la curva centrada de la ecuación:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{X^2}{2\sigma^2}}$$

Para obtener la distribución con la desviación estándar igual a 1, se realiza una segunda transformación que consiste a remplazar ***X*** por ***Z***.

$$Z = X/\sigma$$

Esto conduce a la curva normal centrada y reducida de la forma:

$$f(Z) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{Z^2}{2}} \quad \infty < Z < \infty$$

Si se remplaza ***x*** por ***Z*** con ***Z = x- μ/σ***, se obtiene el mismo resultado, esta ley normal centrada y reducida no es mas que un caso particular de ley normal (figura 15).

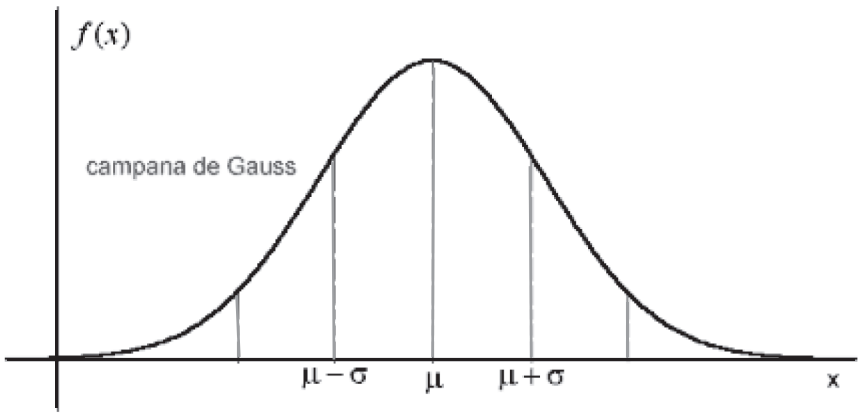


Figura 15. Distribución normal.
Tomado de, http://personal5.iddeo.es/ztt/Tem/t21_distribucion_normal.htm

3.1.3.2. Características de la distribución normal

a) Es simétrica respecto a su media

b) La media, mediana y moda son iguales

c) El área total debajo de la curva por encima del eje ***x*** es una unidad cuadrática en donde el 50% de la información se localiza a la derecha y el 50% a la izquierda.

d) Si se levantan líneas perpendiculares de una distancia estándar de la media, el área encerrada será aproximadamente el 68% del área total. Si se extienden a 2 veces la desviación estándar hacia cada lado de la media se encierra aproximadamente el 95% y hasta 3 veces la desviación del área es de 99% (figura 16).

e) La distribución normal queda determinada por μ y σ , por tanto hay una distribución normal diferente para cada valor diferente de μ y σ de esta manera los valores de μ trasladan a la gráfica de la distribución sobre el eje ***X*** (figura 17), y los valores de σ van a determinar el "aplanamiento" de la curva (figura 18).

3.1.3.3. Propiedades de la distribución normal

a) El fenómeno depende de numerosos factores y estos factores son independientes entre ellos.

b) Los efectos aleatorios de estos factores son acumulativos.

c) Las variaciones de estos factores son bajas y la variación del fenómeno debido a la variación de cada uno de estos factores es también baja.

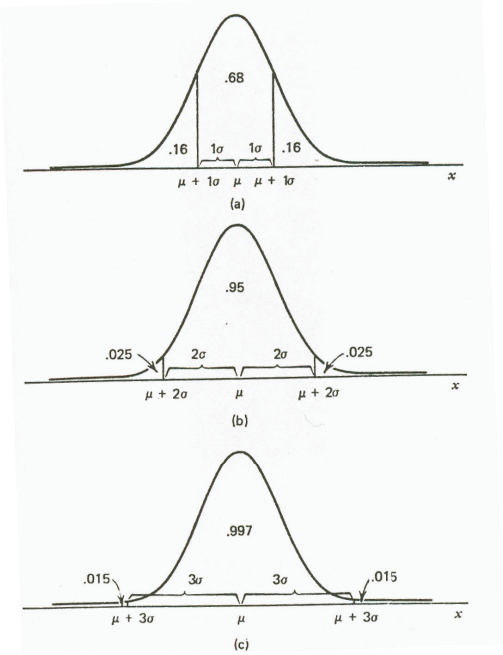


Figura 16. Áreas bajo la curva de la distribución normal.
Tomado de Daniel *et al* (1982), (pág. 78).

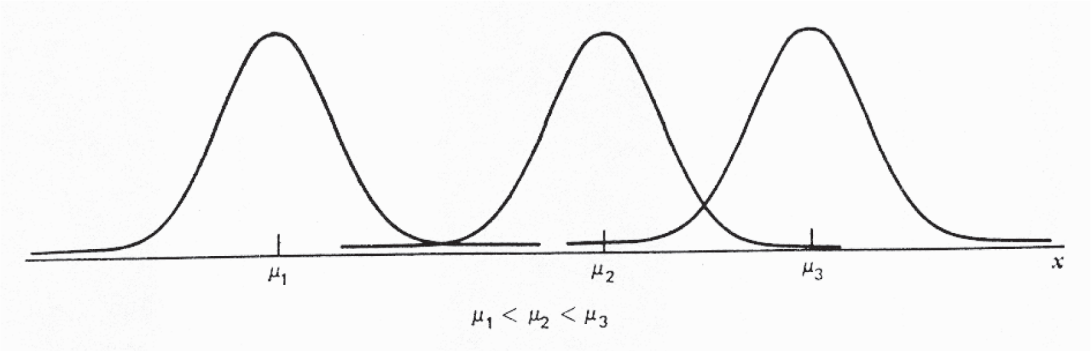


Figura 17. Curvas de distribución normal para diferentes valores de μ.
Tomado de Daniel *et al.* (1982) (pág. 79).

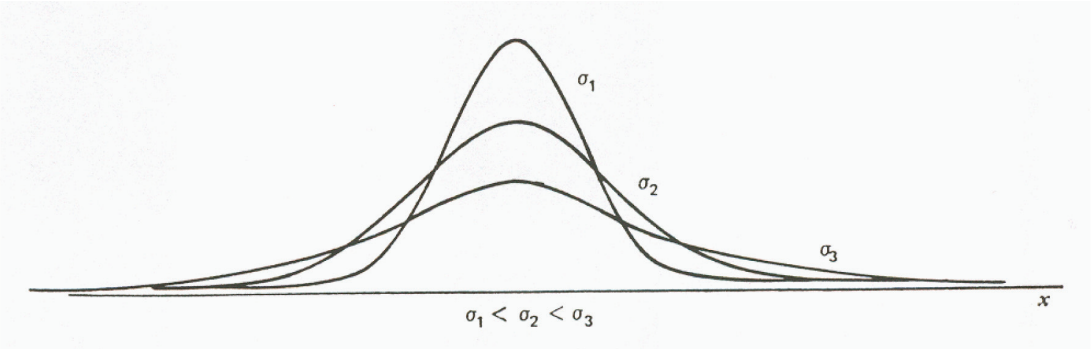


Figura 18. Curvas de distribución normal con diferentes valores de σ.
Tomado de Daniel *et al.* (1982) (pág. 79).

Para encontrar la probabilidad de que ***Z*** tome un valor entre 2 puntos del eje de las ***Z*** (ej. Z_0 y Z_1) se debe encontrar el área limitada entre estos puntos. Esto se obtiene de las tablas de áreas bajo la curva conocidas como tablas de ***Z*** (tabla 3, anexo)

Ejemplo 14. Dada la distribución normal unitaria, encontrar el área bajo la curva por encima del eje ***Z*** entre 0 y ***Z*** = 2

$$P(Z \leq 2) = 0.4772$$

Localizar ***Z*** = 2 en tabla, él área = 0.4772

Esta se puede interpretar entonces como la probabilidad de que ***z*** elegida al azar de las probabilidades de ***Z*** tenga un valor entre 0 y 2 ó también puede interpretarse como la frecuencia relativa de la ocurrencia de los valores de ***Z*** entre 0 y 2.

Ejemplo 15. ¿Cuál es la probabilidad de que una Z elegida al azar tenga un valor entre –2.55 y 2.55?

P(-2.55 < ***Z*** < 2.55) = ***P***(0.4946+0.4946) =0.9892

Las siguientes figuras 19 a-f muestran diferentes formas de calcular las áreas de la curva normal. .unitaria.

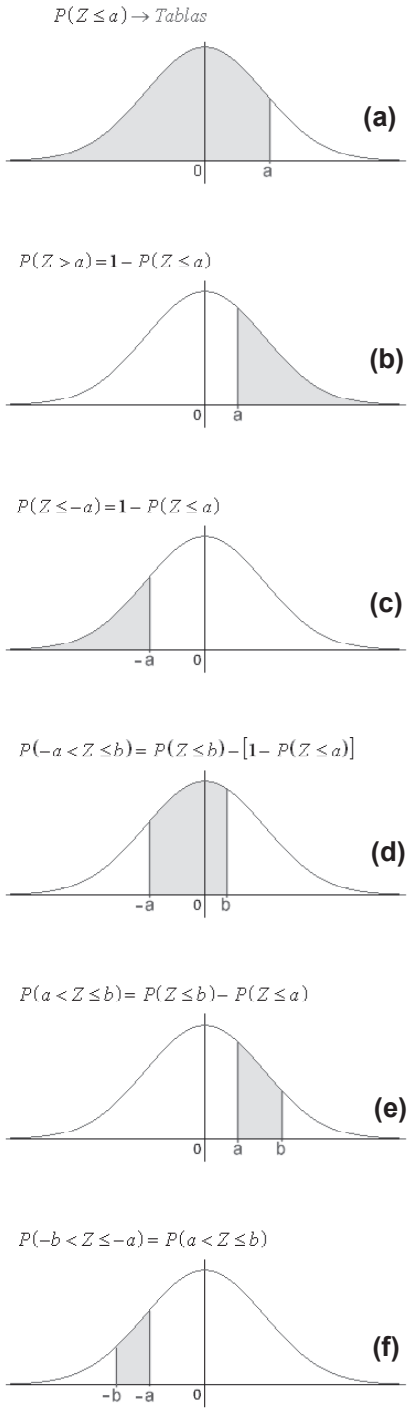


Figura 19. Cálculo de áreas bajo la curva normal unitaria, tomado de http://personal5.iddeo.es/ztt/Tem/t21_distribucion_normal.htm

3.1.3.4. Los momentos de la distribución normal

1. La esperanza matemática $E(X) = \mu$ si $Z = X-\mu/\sigma$

   por lo tanto $X = \sigma Z + \mu$, como $E(X) = E(\sigma z + \mu) = \sigma E(Z) + \mu$

   entonces $E(X) = 0$ y

   por tanto $E(X) = \mu$.

2. La varianza. $Var(X) = \sigma^2$

3. El coeficiente de asimetría, $\alpha_{3=0}$

4. El coeficiente de curtosis, $\alpha_{4=3}$

3.1.4. La Distribución *t-student*

Las desviaciones de las medias muestrales con respecto a la media poblacional presentan comúnmente una distribución normal. Si estas desviaciones son divididas entre la desviación estándar de la población, dichas desviaciones aún estarán distribuidas de acuerdo con la distribución gausiana, con $\mu = 0$, $\sigma = 1$.

Si por otra parte se calcula la varianza s^2, de cada una de las muestras, y la desviación de cada media, como $(x_i - \mu)/s_x$, donde s_x significa la estimación del error estándar de la media de la ***i-ésima*** muestra; el cálculo refleja la distribución de las desviaciones, las cuales pueden variar respecto a la distribución normal. La nueva distribución presenta una forma más ancha que su correspondiente distribución normal, debido a que el denominador representa el error estándar de la muestra en vez del error estándar de la población, por lo tanto esta relación podría resultar en mayores valores de varianza a partir de la relación $(x_i - \mu)/s_x$. La distribución esperada de esta relación, se conoce como distribución *t-student*, planteada por vez primera por el estadístico inglés William Sealy Goset (1876-1937) quien utilizó el seudónimo de “Student” para publicar algunos de sus notables trabajos desarrollados sobre estadística teórica y práctica.

Como la distribución normal, la distribución ***t***, es simétrica y se extiende de manera negativa y positiva hasta infinito. La distribución ***t***, difiere de la distribución normal en que presenta diferentes formas dependiendo del número de grados de libertad. Los grados de libertad $(n-1)$ son el divisor mediante el cual se obtiene una estimación insesgada de la varianza de una suma de cuadrados.

Los grados de libertad pueden variar de 1 a ∞. Una distribución ***t*** para grados de libertad igual a uno, representa la mayor desviación de la forma de una distribución normal. Conforme los grados de libertad se incrementan, la distribución ***t*** se aproxima a la forma de una distribución normal ($\mu = 0$, $\sigma = 1$) y al contrastar los gráficos de una distribución ***t-student*** con 30 grados de libertad con respecto a una distribución normal, la diferencia es esencialmente indistinguible (figura 20).

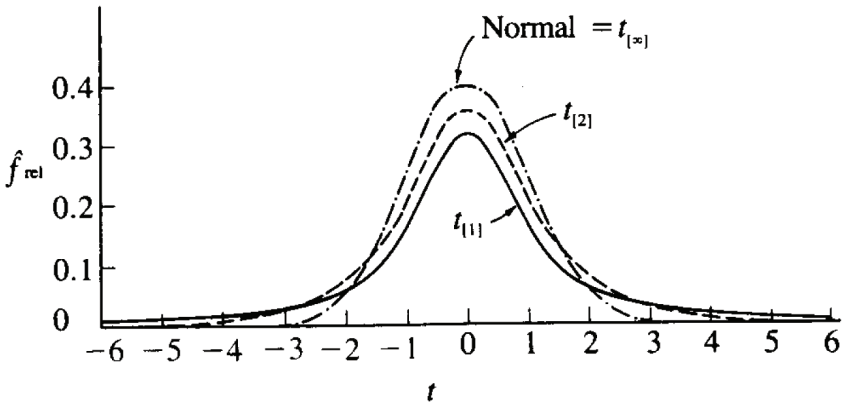


Figura 20. Curvas de frecuencias de distribuciones t (ver tabla 4 anexa) para 1 y 2 grados de libertad, comparadas con la distribución normal, tomado de Sokal y Rohlf (2000) (pag 145).

Con grados de libertad ***(gl)*** igual a ∞ la distribución ***t*** es igual a la distribución normal.

De esta forma podríamos tomar a la distribución ***t*** como un caso general de la distribución normal donde ***gl* = ∞**.

Así debido a que la distribución ***t*** difiere de manera significativa de la distribución normal cuando ***n=30***, es recomendable aproximar una distribución de frecuencias a la distribución ***t-Student***, cuando el tamaño de la muestra es menor a treinta datos.

La ley de ***t-student*** se utiliza en test de comparación de parámetros como la media y en la estimación de parámetros de la población a partir de la información que da la muestra. De esta manera con una tabla de ***t*** se obtienen las probabilidades de obtener un valor de ***t*** al exterior del intervalo ***(-t∞/2 y t∞/2)***; ***P|t| > t∞ / 2) =∞*** (ver tabla 4 anexa)

Así por ejemplo: si ***gl*** = 8 y ∞ = 0.05, |***t***∞/2 = 2.306|

si ***gl*** = 10 y ***α*** = 0.01 |***t***∞/2= 3.170|

3.1.4.1 Momentos de la distribución *t-student*

Su ***Mo*** = 0	***Mo*** = moda
Su ***Me*** = 0	***Me*** = mediana
E(t) = 0	si ***gl*** > 1
Var (t) = gl/gl-2	si ***gl*** > 2

3.1.5 La Distribución Chi-cuadrada χ^2

La distribución χ^2 al igual que la ***t-student*** y la normal, son fundamentales en estadística inferencial, específicamente la distribución χ^2 en conexión con los límites de confianza y distribución de varianzas.

La distribución χ^2 es una función de densidad de probabilidad en la cual los valores van de cero hasta infinito. Así como la distribución normal o ***t***, la función aproxima el eje χ^2 asintóticamente hacia el lado derecho de la curva y no hacia ambos lados.

Al igual que la distribución ***t***, no hay únicamente una sola distribución χ^2, sino una distribución χ^2 para cada número de grados de libertad.

Por lo tanto, la distribución χ^2 es una función de los grados de libertad como se muestra en la figura 21. Como se observa la distribución χ^2 tiende a la simetría a medida que los grados de libertad aumentan.

La distribución χ^2 puede ser generada a partir de una población con desviaciones estándar normales con base en valores normalizados o estandarizados ($(x_i - \mu)/\sigma$); con repetidas muestras de tamaño *n*, y estandarizadas con la expresión previamente señalada.

Las cantidades $\sum^n Z_i^2$, calculadas para cada muestra estarán distribuidas como una distribución χ^2 con ***n*** grados de libertad. Con los datos normalizados podemos escribir $\sum^n Z_i^2$ como:

$$\sum^n Z_i = \frac{\sum_{i=1}^{n}(xi - \mu)^2}{\sigma^2}$$

Después de muestreos repetidos o sucesivos y la estimación de estadísticos, se puede obtener una distribución de frecuencias para esos estadísticos y calcular sus distribuciones estándar como en el caso de una distribución de frecuencia de la media. En algunos casos esos estadísticos estarán distribuidos normalmente como en el caso de las medias. En otros casos los estadísticos estarán distribuidos normalmente sólo si las muestras provienen de una población normal o de muestras lo suficientemente grandes.

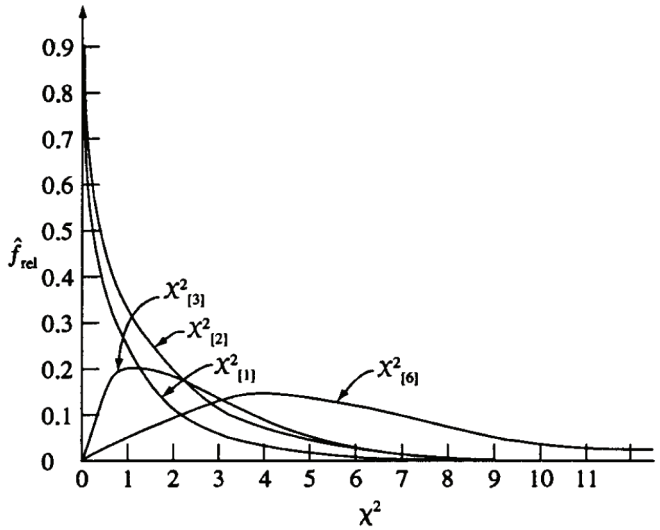


Figura 21. Curvas de frecuencias de distribuciones de χ^2 para 1, 2, 3 y 6 grados de libertad. Tomado de Sokal y Rohlf (2000) (Pag. 153).

En la figura 22 se muestra una distribución de frecuencias de las varianzas de 1400 muestras de 6 datos referentes a las longitudes de ala de moscas.

Es importante destacar que la distribución se encuentra fuertemente sesgada hacia la derecha, lo cual es una característica de la distribución de varianzas obtenidas a partir de muestras pequeñas.

En la figura 22, también se muestran las frecuencias absolutas esperadas de una distribución χ^2, con 4 grados de libertad, la cual coincide bien con la distribución de frecuencias de las varianzas observadas directamente de los datos. Como se mencionó anteriormente, existen tablas χ^2 (ver tabla 5 anexa) para diferentes funciones de repartición de acuerdo a los grados de libertad, aquí los valores de χ^2 corresponden a las probabilidades $\alpha = P(\chi^2 > \chi^2_\alpha)$, por ejemplo el valor de χ^2_α con 12 gl $\chi^2_{(12)}$ a un nivel de $\alpha_{0.5}$ = 21.03.

Aplicaciones de χ^2:

a) en test de comparación de proporciones

b) en test de independencia de 2 caracteres cualitativos

c) en test de comparación de medias experimentales

d) intervalos de confianza de una varianza

e) en test de comparación de una varianza experimental

f) en test no paramétricos.

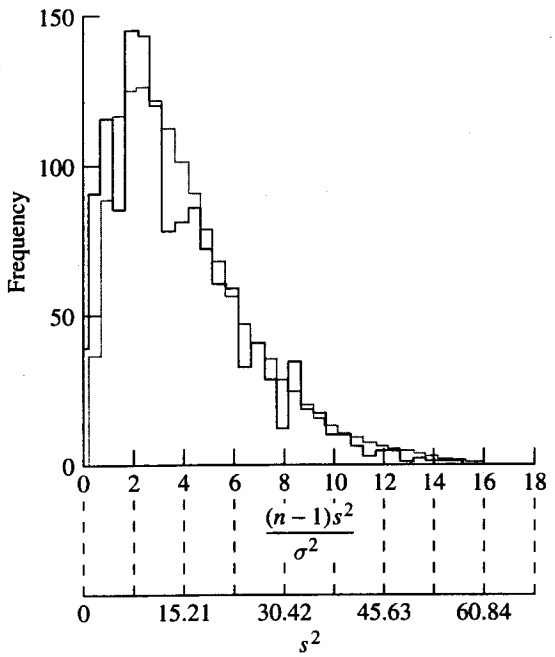


Figura 22. Histogramas de varianzas y valores esperados de la distribución χ^2 con 4 grados de libertad, tomado de Sokal y Rohlf (2000) (Pag. 136).

3.1.5.1 Momentos de la distribución χ^2

1. ***Mo*** = ***gl***-2 (moda) si gl> 2

2. ***E***(χ^2) = ***gl*** (esperanza matemática)

3. ***Var*** (χ^2) = 2***gl*** (varianza)

3.1.6 La Distribución *F* de Fisher

Definición: si $\chi^2_{1\,y}\,\chi^2_{2}$ son una pareja de variables aleatorias independientes que siguen respectivamente la ley χ^2 con gl_1 y gl_2, entonces:

$$F = \frac{\chi_1^2 / gl_1}{\chi_2^2 / gl_2}$$

La función ***f(F)*** depende de gl_1 y gl_2, por lo tanto existen tantas curvas de densidad de probabilidad que valores de gl_1 y gl_2. Así para identificar cada una de las funciones se utilizan las tablas de ***F***.

$$\alpha = P(F(gl_1, gl_2) > F\,\alpha(gl_1, gl_2))$$

Para describir la distribución ***F***, se podría plantear un ejemplo donde se supone realizar un muestreo al azar de una población normalmente distribuida. El procedimiento de muestreo consiste en tomar una primera muestra n_1 y calcular su varianza s_1^2, seguida de una segunda muestra n_2 con varianza s_2^2. Las muestras n_1 y n_2 posiblemente sean del mismo tamaño o no.

Así por ejemplo, si se dividen las varianzas de ambas muestras, mediante la relación, $F_s = s_1^2 / s_2^2$; el cociente estará muy cerca a la unidad, porque las varianzas son estimadas de la misma población.

Si se toman muestras repetidas de tamaños n_1 y n_2, y se calculan las relaciones F_s de sus varianzas el promedio de estas relaciones se aproximará a $((n_2 - 1)/(n_2 - 3))$, la cual es cercana a 1.0 cuando n_2 es grande. Esta distribución se conoce como distribución ***F***, en honor a Ronald Aylmer Fisher.

Las relaciones entre varianzas muestrales (s_1^2 / s_2^2) son estadísticos de las muestras que posiblemente sigan una distribución ***F***, dependiendo si las muestras siguen una distribución normal.

La figura 23 muestra diferentes distribuciones ***F*** representativas. Para valores bajos de grados de libertad la forma de la distribución ***F*** se aproxima a una ***"L"***, sin embargo a medida que se incrementan los grados de libertad se forma un sesgo más pronunciado hacia la derecha.

Los valores en una tabla $F_{\alpha,gl1,gl2}$ (tabla 6 anexa), incluyen los valores de α, la cual es la proporción de área bajo la curva del lado derecho, con gl_1 (numerador) y gl_2 (denominador) de la relación entre varianzas respectivamente. La tabla de las proporciones de la distribución ***F***, se encuentra ordenada con gl_1 que corresponde a la varianza superior (numerador) entre las dos muestras, dispuesta a lo largo del margen superior de la tabla y los gl_2 que corresponde a la varianza menor (denominador), cuyos valores se encuentran a lo largo del margen lateral de la tabla de la distribución ***F***.

Tres distribuciones *F* representativas

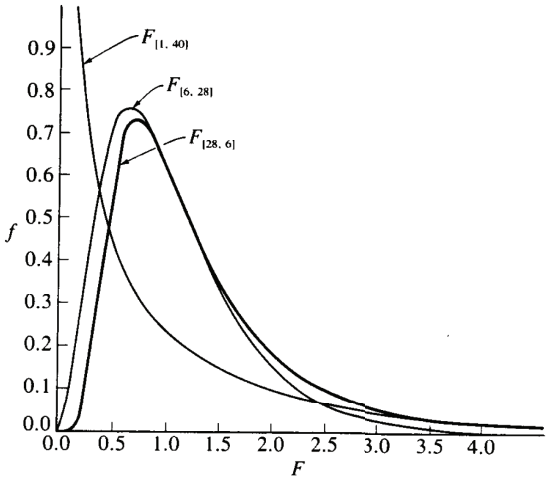


Figura 23. Curvas de distribución de F para diferentes grados de libertad.
Tomado de Sokal y Rohlf (2000) (pag. 186).

Dado que en la pruebas de hipótesis usualmente se toma interés en valores bajos de ***α***, el extremo derecho de la curva es especialmente enfatizado en las tablas ***F***.

Así por ejemplo una distribución ***F*** con $gl_1 = 6$, $gl_2 = 28$ es de 2.45 a $\alpha = 0.05$; esto indica que 0.05 del área bajo la curva se presenta hacia el lado derecho de ***F*** = 2.45 (figura 24).

Dado a que es común utilizar solo el lado de la derecha de una distribución ***F***, cuando se plantea una prueba de hipótesis de dos extremos es necesario obtener valores de alfa > 0.5 (lo cual corresponde al lado izquierdo de la distribución ***F***) por lo tanto es necesario aplicar la siguiente relación:

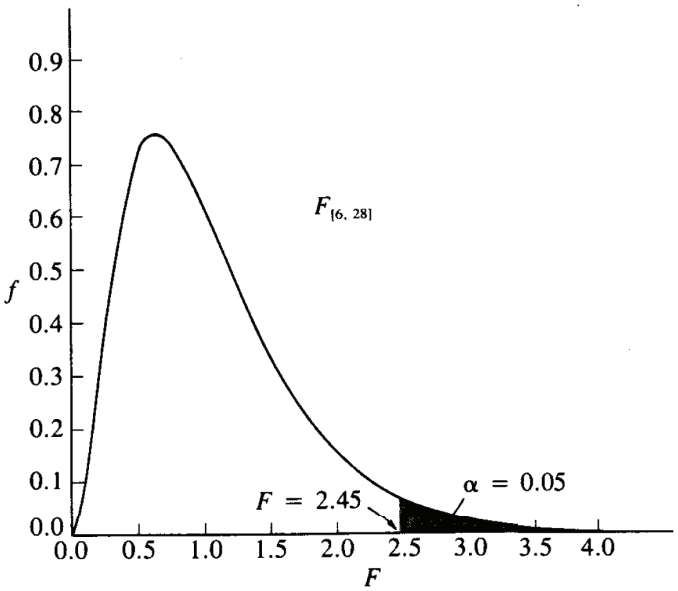


Figura 24. Distribución F con $gl_1 = 6$, $gl_2 = 28$; F = 2.45 a $\alpha = 0.05$.
Tomado de Sokal y Rohlf (2000) (Pag. 187).

$$F_{\alpha, gl1, gl2} = \frac{1}{F_{1-\alpha, gl2, gl1}}$$

Es importante destacar con base en la expresión de arriba que si $F_{0.05,5,24} = 2.64$, y deseamos obtener $F_{0.95,5,24}$; debemos encontrar el recíproco de $F_{0.05,24,5} = 4.56$.

3.1.6.1 Momentos de la distribución F

a) la esperanza matemática

$$E(F) = \frac{gl_2}{gl_2 - 2} \qquad si\ gl_2 > 2$$

b) varianza

$$Var(F) = \frac{2gl_2^2(gl_1 + gl_2 - 2)}{gl_1(gl_2 - 2)^2 (gl_2 - 4)} \qquad si\ gl_2 > 4$$

Aplicaciones de ***F***:

La variable aleatoria ***F*** se utiliza para comparar 2 varianzas experimentales y sirve en test de análisis de varianza y covarianza. La covarianza es una generalización del concepto de varianza en un espacio de dos dimensiones (figura 25).

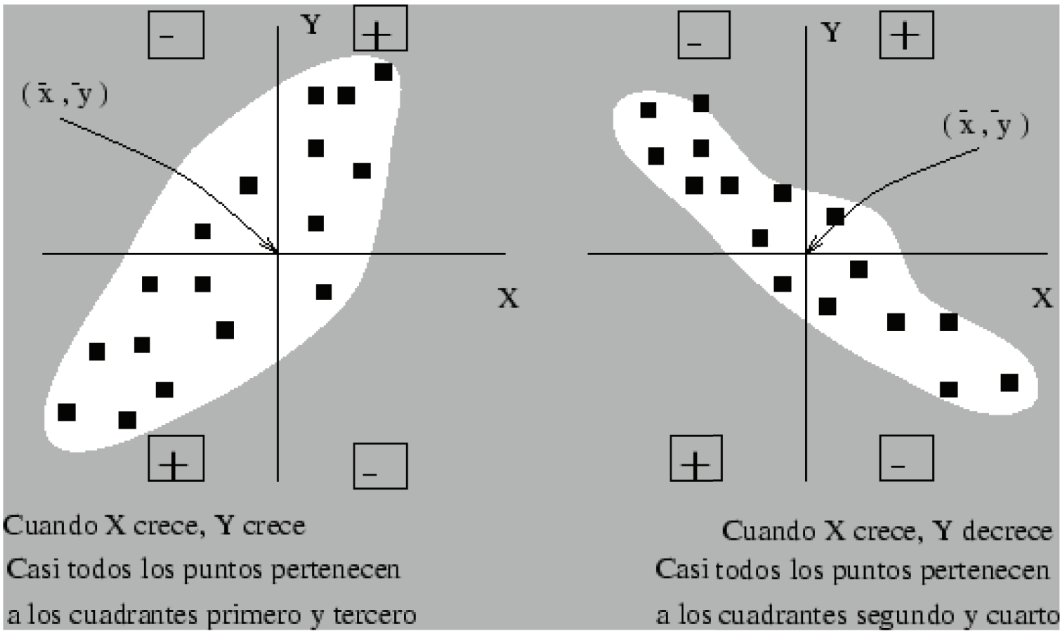


Figura 25. Interpretación geométrica de $S_{XY,}$
tomado de http://www.bioestadistica.uma.es/libro/node38.htm

$$S_{xy} = \frac{\sum_{i=1}^{N} (X_i - \bar{X})\, Y_i - \bar{Y}}{N} \qquad \text{para poblaciones; y}$$

$$S_{xy} = \frac{\sum_{i=1}^{N} (x_i - \bar{x})\, y_i - \bar{y}}{N - 1} \qquad \text{para muestra}$$

Así se deduce la varianza cuando $\boldsymbol{x = y}$

$$\boldsymbol{S} = \begin{bmatrix} Sx^2 & S_{XY} \\ S_{yx} & Sy^2 \end{bmatrix}$$

De este modo:

- Si hay mayoría de puntos en el tercer y primer cuadrante, ocurrirá que $\boldsymbol{S_{XY} \geq 0}$, lo que se puede interpretar como que la variable $\boldsymbol{Y}$ tiende a aumentar cuando lo hace $\boldsymbol{X}$;
- Si la mayoría de puntos están repartidos entre el segundo y cuarto cuadrante entonces $\boldsymbol{S_{XY} \leq 0}$, es decir, las observaciones $\boldsymbol{Y}$ tienen tendencia a disminuir cuando las de $\boldsymbol{X}$ aumentan;
- Si los puntos se reparten con igual intensidad alrededor de $(\bar{\boldsymbol{x}}, \bar{\boldsymbol{y}})$, entonces se tendrá que: $\boldsymbol{S_{XY} = 0}$.

Cuando los puntos se reparten de modo más o menos homogéneo entre los cuadrantes primero y tercero, y segundo y cuarto, se tiene que $\boldsymbol{S_{XY} \approx 0}$ (figura 26). Eso no quiere decir de ningún modo que no pueda existir ninguna relación entre las dos variables, ya que ésta puede existir como se aprecia en la figura 26 (derecha).

Para concluir, en la covarianza se tiene que:

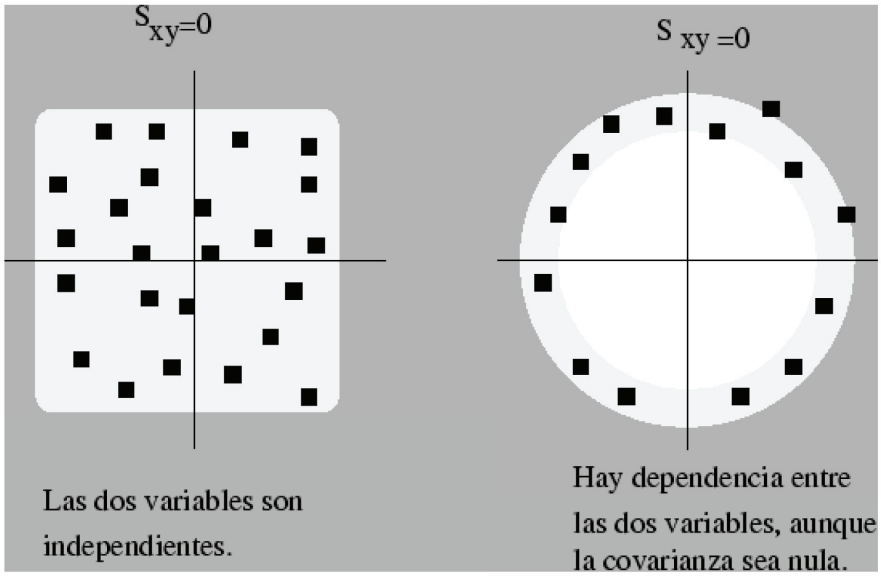


Figura 26. Repartición homogénea de los puntos cuando $S_{XY} \approx 0$.
Tomado de http://www.bioestadistica.uma.es/libro/node38.htm

• Si $S_{XY} > 0$, las dos variables crecen o decrecen a la vez (nube de puntos creciente).

• Si $S_{XY} < 0$ cuando una variable crece, la otra tiene tendencia a decrecer (nube de puntos decreciente).

• Si los puntos se reparten con igual intensidad alrededor de $(\bar{x}, \bar{y})$, $S_{XY} = 0$, (no hay relación lineal).

3.2 La Inferencia Estadística

Literatura sugerida:

Daniel.W. W., 1982. Biostadística. Limusa. Mexico. 485 p (pág. 1, 140, 142)

Scherrer B., 1984. Biostatistique. Gaëtan Morin Editor. Montreal, Paris, Casa Blanca. 850 p (Pág 337-343).

Zar. J. H., 1999. Bostatistical Analysis (4 edición). Prentice Hall. Estados Unidos. 663 p. (Pag. 91-92, 122-131).

http://www.itcomitan.edu.mx/tutoriales/estadistica/contenido/unidad_4.html

La inferencia estadística de acuerdo a Daniel, 1982 es el procedimiento por el que se llega a inferencias respecto a una población, con base en los resultados que se obtienen a partir de una muestra extraída de esa población. La inferencia estadística incluye dos áreas generales: la **estimación** y las **pruebas de hipótesis**.

La **estimación** conlleva a calcular, basado en los datos de una muestra, alguna estadística que puede ser vista como una aproximación del parámetro correspondiente a la población, de la cual se extrajo la muestra. De esta manera se habla de **estimación puntual,** en el caso de por ejemplo, el valor promedio de la característica (variable) que se este midiendo y de **estimación de intervalo** al conjunto de dos números entre los cuales se pretende que se encuentre la cantidad desconocida. Al dar una estimación de intervalo suele proporcionarse una medida de la **confianza** que se tiene de la estimación. La regla o fórmula que indica como calcular el valor de la estimación se le conoce como **estimador** así por ejemplo:

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

Es un estimador de la media poblacional μ y el valor único que resulta de la evaluación de esta fórmula se llama estimación del parámetro μ.

Las **pruebas de hipótesis** se basan en decisiones estadísticas. En la práctica el investigador debe de tomar ciertas decisiones sobre la población a partir de la información proporcionada por una o muchas muestras. Así por ejemplo el decidir sobre las hipótesis de si un nuevo medicamento es efectivamente eficaz para curar una enfermedad, o si la densidad de una especie animal es igual de alta en dos medios de cultivo, o si la concentración de DDT en el tejido adiposo no aumenta con la edad de los sujetos, tales decisiones se llaman decisiones estadísticas. Más adelante se hablará más a fondo sobre estas hipótesis y a continuación se señalarán las estimaciones por intervalo o intervalos de confianza, se explica como calcularlos y como se utilizan las pruebas de hipótesis:

3.2.1 Estimación por Intervalo (Intervalos de Confianza)

Las estimaciones dadas por una muestra, no son interpretables que acompañadas de indicadores cuantificables, fijando un grado de confianza que se les pueda otorgar. Un intervalo de confianza indica la prescisión de una estimación, puesto que para un riesgo de α dado, el intervalo es más grande que la prescisión es menor. La estimación por intervalo se realiza cuando algún parámetro de la población se encuentra ubicado dentro de un límite inferior y un límite superior según el nivel de confiabilidad a considerar, donde la población que se analiza se distribuye de acuerdo a una distribución normal. El resultado final está expresado en términos de la probabilidad de que un parámetro determinado se ubique dentro de cierto intervalo de valores, bajo un grado de confianza definido.

Para llevar a cabo la estimación por intervalo de algún parámetro de la población es necesario partir de los estadísticos o estimadores, apoyados de las proporciones de distribución continuas, tomadas de las tablas correspondientes (***t-student***, χ^2 y ***Z***).

3.2.1.1 Estimación por intervalo para la media poblacional

La estimación por intervalo con $\alpha = 0.05$, para la media se basa en la siguiente expresión:

$$P(\bar{X} - t_{\alpha/2}\, S\bar{x} < \mu < \bar{X} + t_{\alpha/2}\, S\bar{x}) = 1 - \alpha \qquad \text{para } \boldsymbol{n} < 30$$

$$P(\bar{X} - Z_{\alpha/2}\, S\bar{x} < \mu < \bar{X} + Z_{\alpha/2}\, S\bar{x}) = 1 - \alpha \qquad \text{para } \boldsymbol{n} > 30$$

El error tipo será para ambas estimaciones:

$$S_{\bar{x}} = \frac{S_x}{\sqrt{n-1}} \qquad \text{para } \boldsymbol{t} \text{ y } \boldsymbol{v} = n-1$$

$$S_{\bar{x}} = \frac{S_x}{\sqrt{n}} \qquad \text{para } \boldsymbol{Z}$$

Si se tiene conocimiento de la población total la ecuación del error tipo se modifica en:

$$S_{\bar{x}} = \frac{S_x}{\sqrt{n}} \sqrt{\frac{N-n}{N}}$$

donde, ***N*** es de la población y ***n*** de la muestra

Donde la primera expresión es utilizada para muestras donde $\boldsymbol{n} \leq 30$, y la segunda expresión cuando $\boldsymbol{n} > 30$, con base en la distribución ***t-student*** y la distribución normal, para la búsqueda de los valores críticos, $\boldsymbol{t}_{\alpha/2,v}$ y $\boldsymbol{z}_{\alpha/2)}$, respectivamente. La expresión, $S_{\bar{x}}$, indica el error estándar asociado a la muestra. (notar que en la expresión $\boldsymbol{t}_{\alpha/2,v}$, la $\boldsymbol{v}$ indica grados de libertad)

La cantidad referida $\boldsymbol{(\bar{X} - t_{\alpha/2}\, S_{\bar{X}})}$, o $\boldsymbol{(\bar{X} - z_{\alpha/2}\, S_{\bar{X}})}$, es llamada limite de confianza inferior (abreviado como $\boldsymbol{L_1}$); y la señalada $\boldsymbol{(\bar{X} + t_{\alpha/2}\, S_{\bar{X}})}$, o $\boldsymbol{(\bar{X} + z_{\alpha/2}\, S_{\bar{X}})}$, es llamado limite de confianza superior (abreviado como $\boldsymbol{L_2}$).

Por definición los coeficientes de confiabilidad al 95% o α= 0.05, al 99% o α= 0.01 y al 99.9% α =0.001son:

$P(\bar{X} - 1.96\, S_{\bar{x}} < \mu < \bar{X} + 1.96\, S_{\bar{x}}) = 0.95$ $\quad \alpha = 0.05$

$P(\bar{X} - 2.57\, S_{\bar{x}} < \mu < \bar{X} + 2.57\, S_{\bar{x}}) = 0.99$ $\quad \alpha = 0.01$

$P(\bar{X} - 3.3\, S_{\bar{x}} < \mu < \bar{X} + 3.3\, S_{\bar{x}}) = 0.99$ $\quad \alpha = 0.001$

Ejemplo 16. Supongamos que necesitamos determinar la estimación por intervalo para la media, μ, de una población que sigue una distribución normal con desviación estándar, S_x = 5.1, en una prueba donde se quiere conocer el promedio de oxígeno disuelto que se tiene en un sistema acuícola, con base en un nivel de confianza de 95%, a partir de una muestra n= 100, con un promedio de 4.2 mg/l.

Para la solución del problema, utilizamos la distribución normal debido a que $n > 30$, obteniendo el estadístico teórico, $Z_{0.05(2)} = 1.96$;

un error estándar, $S_{\bar{x}} = \dfrac{5.1}{\sqrt{100}} = 0.51$

por lo tanto el limite inferior sera, $L_1 = 4.2-[1.96(0.51)] = 3.2$,

y el límite superior sera, $L_2 = 4.2+[1.96(0.51)] = 5.2$.

El resultado se expresa de la siguiente forma:

$$P(3.2 \leq \mu \leq 5.2) = 0.95$$

Esto indica que hay un 95% de posibilidades de que la media del oxígeno disuelto sea superior a 3.22 mg/l e inferior a 5.22 mg/l.

3.2.1.2. Estimación por intervalo para la varianza poblacional

Los intervalos de confianza pueden también ser determinados para otros parámetros de la población, con el propósito de expresar una mayor precisión en las estimaciones de esos parámetros.

La distribución muestral de la media es una distribución simétrica, la cual se aproxima a una distribución normal a medida que el valor de n se incrementa. Pero la distribución muestral de la varianza no es simétrica y ni la distribución normal o la distribución t pueden ser empleados para determinar los limites de confianza alrededor de la varianza poblacional o para probar hipótesis acerca de la varianza poblacional, σ^2. Entonces se debe emplear la distribución χ^2 para definir los intervalos de confianza correspondientes a la varianza de la población.

Si se desean conocer los dos valores de χ^2 que incluye 1 - α de la curva de chi-cuadrada, debemos de encontrar la porción de la curva entre $\chi^2_{(1-\alpha/2),\,v}$ y $\chi^2_{(\alpha/2),\,v}$ (Para un 95% de intervalo de confianza, esto sería entre ($\chi^2_{0.975,\,v}$ y $\chi^2_{0.025,\,v}$). Dichos intervalos de confianza se calculan a través de la siguiente expresión:

$$P\left(\frac{(n-1)\,s_x^2}{\chi^2_{\alpha/2}\ 0.025} \le \sigma^2 \le \frac{(n-1)\,s_x^2}{\chi^2_{\alpha/2}\ 0.975}\right) = 0.95$$

$$P\left(\frac{(n-1)\,s_x^2}{\chi^2_{\alpha/2}\ 0.005} \le \sigma^2 \le \frac{(n-1)\,s_x^2}{\chi^2_{\alpha/2}\ 0.995}\right) = 0.99$$

S^2_x es la varianza muestral y corresponde a los estadísticos teóricos basados en las proporciones de la distribución de χ^2.

Ejemplo 17. En una investigación se pretende analizar la variabilidad del tiempo de reacción de cierto medicamento aplicado a una especie de peces, donde la población sigue una distribución normal con un 5% de error, tomándose una muestra de n = 20 con una varianza muestral s_x^2 = 72.25 hr^2. Cual será la estimación por intervalo para la varianza poblacional.

Para la solución del problema expuesto arriba, como se mencionó previamente, utilizamos la distribución chi-cuadrada, obteniendo los estadísticos teóricos, $\chi^2_{0.975,19}$ (buscado en tabla, para el limite inferior, L_1) y $\chi^2_{0.025,19}$ (buscado en tabla, para el limite superior, L_2); y la varianza muestral, s_x^2 = 75.25, por lo tanto el limite inferior, L_1 = [19(72.25)]/32.85 = 41.78, así mismo el límite superior, L_2 = [19(72.25)]/8.91 = 154.07.

El resultado se expresa de la siguiente forma:

$$P(41.78 \le \sigma^2 \le 154.08) = 0.95$$

Esto indica que hay un 95% de posibilidades de que la media de la variabilidad de tiempo de reacción sea superior a 41.78 hr^2 e inferior a 154.08 hr^2

3.2.1.3. Estimación por intervalo para la desviación estándar

Para calcular los intervalos de confianza de la desviación estándar, simplemente se saca la raíz cuadrada de los intervalos de la varianza

$$P\sqrt{\frac{(n-1)\,S_x^2}{\chi^2_{\alpha/2(0.25)}}} \le \sigma^2 \le \sqrt{\frac{(n-1)\,S^2_{\bar{x}}}{\chi^2_{1-\alpha/2(0.975)}}} = 0.95$$

$$P\sqrt{\frac{(n-1)\,S_x^2}{\chi^2_{\alpha/2(0.005)}}} \le \sigma^2 \le \sqrt{\frac{(n-1)\,S^2_{\bar{x}}}{\chi^2_{1-\alpha/2(0.995)}}} = 0.95$$

3.3. Tamaño o Talla de Muestra

La pregunta de que tan grande tomar la muestra, surge inmediatamente en la planificación de cualquier investigación o experimento. Esta es una cuestión importante y no debe tomarse a la ligera. Tomar una muestra más grande de lo necesario para alcanzar los resultados deseados, es un desperdicio de los recursos, mientras que muestras pequeñas a menudo conducen a resultados sin uso práctico. Entonces, como se puede proceder acerca de la determinación del tamaño de la muestra que se necesita en una situación dada?.

La talla de muestra ***n*** se calcula como:

$$n = \frac{z^2\sigma^2}{d^2} \qquad o \quad \bar{x}\,E = tn_{-1} * \frac{s}{\sqrt{n}} = \sqrt{\frac{t_{n-1}\,S}{\bar{x} * E}} = n$$

Donde Z es el valor del coeficiente de confianza, σ^2 es la varianza y d^2 que es el intervalo de confianza multiplicado por el error estándar elevado al cuadrado.

Ejemplo 18. Una neuróloga del departamento de salud pública, deseando conducir una investigación entre una población de muchachas adolescentes, con el fin de determinar su ingestión diaria promedio de proteínas, está buscando consejo de un bioestadísitico, relativo al tamaño de la muestra que debe tomar.

¿Que procedimiento sigue el estadístico para dar asistencia a la nutrióloga? Antes de que el estadístico pueda ayudar a la nutrióloga, esta debe dar tres detalles de información: el ancho deseado del intervalo de confianza, el nivel de confianza deseado y la magnitud de la varianza de la población. Supongamos que a la neuróloga le gustaría que su estimación estuviera dentro de 5 unidades aproximadamente de la verdadera en cualquier dirección (intervalo de confianza). Supongamos también que se decide por un coeficiente de confianza de 0.95 (***Z***) y que, de su experiencia pasada, la nutrióloga siente que la desviación estándar de la población es probablemente de alrededor de 20 gramos. Ahora el estadístico tiene la información necesaria para calcular el tamaño de muestra: $\boldsymbol{z} = 1.96$, $\boldsymbol{\sigma^2}$ es 20 y $\boldsymbol{d^2} = 5$. Supongamos que la población de interés es grande, de modo que el estadístico puede ignorar la corrección de población finita y usar la ecuación anterior. El estadístico recomienda a la neuróloga que su tamaño de muestra debe ser de 61 personas

$$n = \frac{(1.96^2)(20^2)}{5^2} = 61.44$$

Cuando el muestreo es sin reemplazo a partir de una población finita, se requiere la corrección por población finita:

$$n = \frac{(Nz^2)(\sigma^2)}{d^2\,(N-1) + z^2\sigma^2}$$

Es importante mencionar que para poblaciones pequeñas, se sustituye ***Z*** por el valor de ***t-student*** al nivel de confianza deseado.

3.4. Pruebas de Hipótesis

Uno de los principales propósitos del análisis estadístico es hacer inferencias acerca de la población mediante la comparación de una o más muestras de la población, para esto hay que formularse hipótesis estadísticas. La primera hipótesis se denomina **hipótesis nula** (Fisher, 1935) abreviada H_0 (Pearson, 1947) esta supone un efecto dado: por ejemplo, podríamos tratar de probar la hipótesis de H_0: **El poder aglutinante del suero es el mismo en los cirróticos que en los sujetos normales, o bien,** H_0: **La densidad del pájaro** χ **es la misma en los bosques tropicales que el los bosques de coníferas**. Esta hipótesis expresa el concepto de no diferencia. Estas hipótesis difieren de la **hipótesis alternativa o contraria (**H_1 **o** H_A**)** y supone lo contrario de H_0. Por ejemplo, H_1**: El poder aglutinante del suero en los cirróticos difiere de los sujetos normales.** H_1: La densidad del pájaro "χ" **es diferente en los bosques tropicales que el los bosques de coníferas**.

Una vez definidas las hipótesis estadísticas, estas deben ser sometidas a una prueba de verdad, es decir definir si los resultados obtenidos por la muestra son conforme a los resultados supuestos, de esta manera definir si las diferencias registradas entre las observaciones y las hipótesis provienen del azar.

3.4.1 Riesgos de Error en un Test Estadístico

Cada test estadístico comporta dos riesgos de error. a) Si se rechaza una hipótesis que debería ser aceptada, se comete un error tipo α, también llamado de primera especie (Tipo I), y por el contrario b) Si se acepta una hipótesis que debería ser rechazada se comete un error de tipo β o de segunda especie (Tipo II). En el ejemplo de los cirróticos, α representa el riesgo de declarar diferente la media del poder aglutinante del suero de los cirróticos y los sujetos normales, mientras que son idénticas a nivel de poblaciones, y β es el riesgo de declarar idénticas las medias de ambos grupos mientras que son diferentes.

Los riesgos de error en un test estadístico

Hipótesis	H_0 es verdad	H_1 es verdad
H_0 aceptada	Buena decisión	Error β
H_0 rechazada	Error α	Buena decisión

El cálculo del riesgo de error, depende de la manera como ha sido construido el test estadístico; así por ejemplo en el caso de formular un test para dimorfismo sexual en camarones a partir de la longitud total, se tiene una muestra de 121 individuos de los cuales se determinó el sexo por observación del télico. La longitud total media de 65 machos es 61.16 mm y la desviación estándar de 1.11 mm. Si se admite que la distribución de la variable es normal, y que la media de la muestra es una buena estimación de la media poblacional, se puede decir que:

$$P(X \geq \bar{X} + Z_\alpha S_x) = 0.05$$

$$P(X \geq 61.16 + 1.64 * 1.11) = 0.05$$

1.64 corresponde al valor de Z para un test unilateral donde $\alpha = 0.05$.

$$P(X \geq 62.98) = 0.05$$

Hipótesis establecidas:

H_0: el camarón extraído aleatoriamente de la población que posee una longitud total X_i es un macho

H_1: el camarón extraído aleatoriamente de la población que posee una longitud total X_i es un hembra

Las reglas de decisión pueden establecerse de la siguiente manera:

- Si X_i es superior al valor crítico 62.98 mm la hipótesis es rechazada y el camarón es considerado como una hembra
- Si X_i es inferior a 62.98 mm la hipótesis principal es aceptada y el camarón es un macho (figura 27).

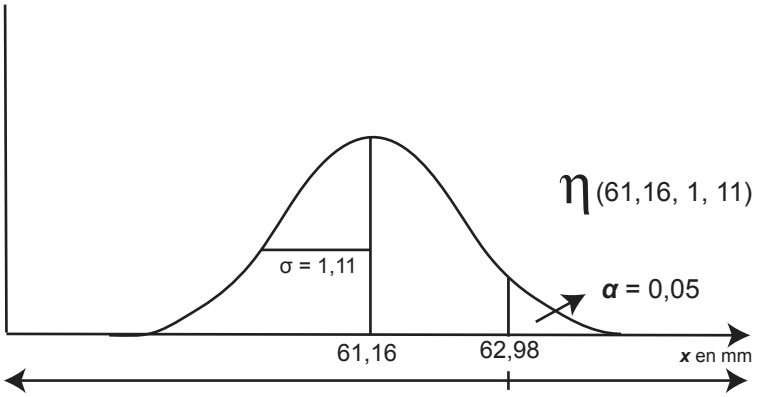


Figura 27. Distribución de la longitud de camarón (mm) y nivel de significativdad. Modificado se Scherrer (1984) (pág, 370)

Con esta regla, el riesgo de tomar a una hembra por un macho es α = 0.05, en cuanto al riesgo β de tomar un macho por una hembra, este puede ser calculado a partir de las características de la población de machos inmaduros.

Si se considera que la longitud media de los camarones hembras, calculada sobre 56 individuos es 63.74 mm con una desviación estándar de 1.20 mm y se admite que la distribución de la variable obedece a una distribución normal, es posible calcular la probabilidad de tener un macho con longitud inferior a 62.98 mm, nivel α = 0.05 (figura 28)

$$P(X < 62.98) = \beta$$

$$P(Z > \frac{62.98 - 63.74}{1.2}) = \beta$$

$$P(X < - 0.716) = \beta$$

Ahora se lee 0.716 en la tabla de áreas bajo la curva de Z, para este valor Z = 0.7611 para obtener β **se** resta 1 al valor obtenido (puesto que el área total bajo la curva vale 1) con lo que β = 0.239 es decir que la probabilidad de tomar a una hembra por un macho es 23.9%.

Para que un test estadístico sea eficaz, debe de estar construido de manera que los errores en la decisición sean mínimos. Lo que no es simple, puesto que para una muestra de un tamaño dado, la disminución de un tipo de error, es normalmente acompañada por el incremento del otro tipo. En la práctica, un tipo de error puede ser más importante que otro, el investigador tiene que definir el nivel de error que arriesga, con el fin de limitar el error más importante.

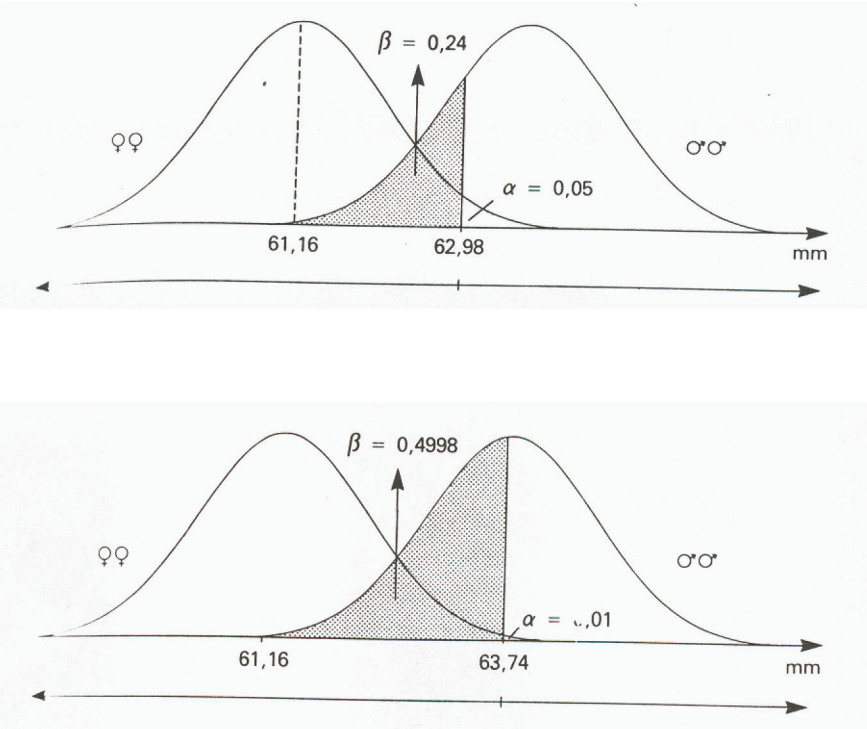


Figura 28. Distribución de la longitud de camarón (mm) y nivel de significatividad y error β. Tomado se Scherrer (1984) (Pág, 371)

3.4.2. Umbral de Probabilidad o Nivel de Significatividad

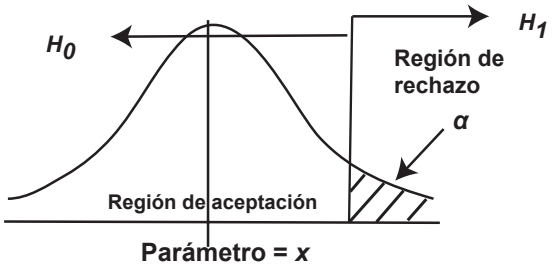
Cuando se prueba una hipótesis se arriesga un error ***α*** hasta un cierto nivel, a este se le llama umbral de probabilidad o nivel de significatividad. Se reconoce como significativo al nivel de probabilidad igual al 0.05, muy significativo en nivel 0.01 y altamente significativo el nivel 0.001. Así si por ejemplo si se escoge el nivel de 0.05 entonces se dice que hay 5 oportunidades sobre 100 de rechazar una hipótesis H_0 cuando esta debe ser aceptada.

3.5 Pruebas de Hipótesis de una Cola o Dos Colas, Llamado también Test Unilateral o Bilateral

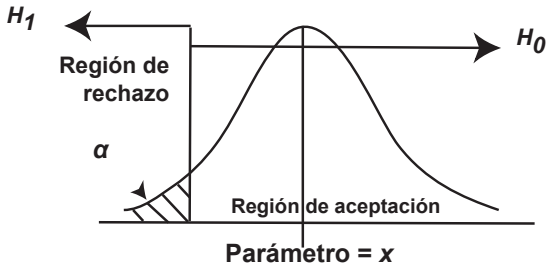
Antes de aplicar todo test estadístico hay que definir el problema propuesto, así según las hipótesis formuladas, se habla de un test bilateral o unilateral. El test bilateral se utiliza cuando se tiene que definir entre dos estimaciones, o entre una estimación y un valor dado sin tener en cuenta el signo o sentido de la diferencia, es decir sólo se utiliza si la H_0 (hipótesis nula) señala la igualdad. El test unilateral se aplica cuando se necesita saber si una estimación H_1 (hipótesis alternativa) es superior o inferior a otra.

Las determinaciones de las zonas de aceptación o de rechazo en un test unilateral y bilateral se muestran en la figura 29:

$H_1 : \mu_1 < \mu_2$ **unilateral**



$H_1 : \mu_1 > \mu_2$ **unilateral**



$H_1 : \mu_1 \neq \mu_2$ **bilateral**

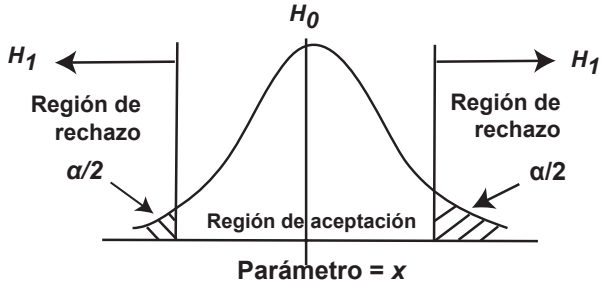


Figura 29. Regiones de aceptación y rechazo de la hipótesis nula para test unilateral y bilateral. Tomado de http://www.itcomitan.edu.mx/tutoriales/estadistica/contenido/unidad_4.html

3.5.1. Pruebas de Hipótesis Paramétricas con una Muestra

Dentro de las pruebas de hipótesis podemos distinguir dos grupos principales, aquellas que trabajan con datos presentados en escala cardinal, donde las muestras siguen una distribución normal, y las varianzas son homogéneas; denominadas **pruebas estadísticas paramétricas**, ya que utilizan los parámetros para hacer las inferencias estadísticas. Por otro lado tenemos las **pruebas estadísticas no paramétricas** que comúnmente se basan en suma de rangos para establecer diferencias entre dos o más muestras. Asimismo las pruebas de hipótesis no paramétricas son utilizadas más como pruebas alternativas debido a su bajo poder estadístico.

En general para plantear una prueba de hipótesis, esta deberá estar basada en algún parámetro de la población (varianza, media, proporción, etc.) y de acuerdo a la muestra o número de pruebas que se pretende comparar.

En la figura 30 se muestra la utilización del test de Z para comparar una media de una muestra observada a una media hipotética o histórica de una población.

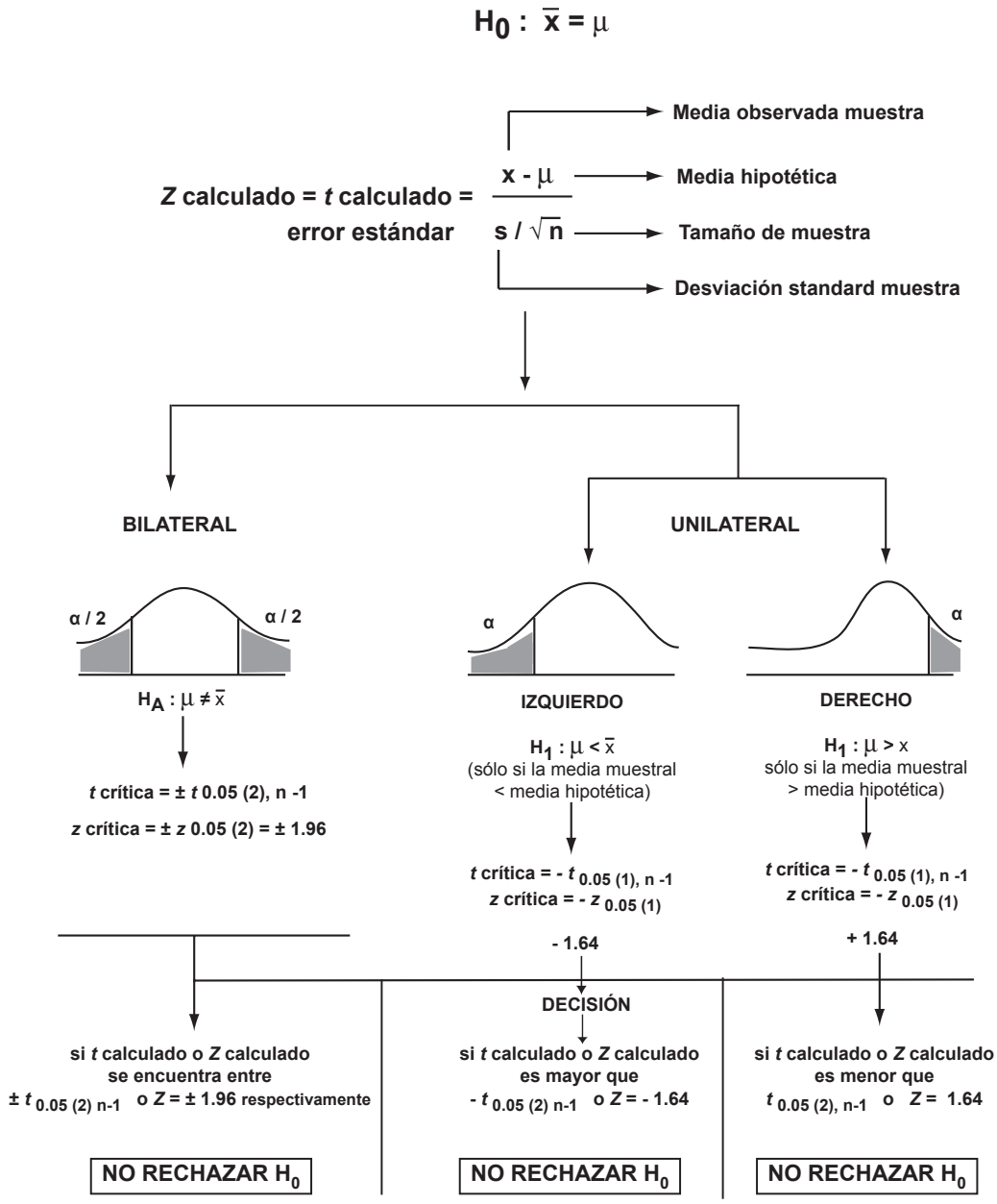


Figura 30. Prueba de hipótesis paramétrica de una muestra (media obervada *vs* hipotética o histórica) H_0: $\mu = \bar{x}$

Ejemplo 19. Se compara la media muestral (10.43 mg/m^3), con respecto a una media poblacional hipotética (10.0 mg/m^3), de 31 datos correspondientes a concentraciones de monóxido de carbono en aire, con un error estandard de 0.24 mg/m^3; no encontrándose diferencias significativas entre dichas medias ($p > 0.05$), es decir a un nivel de significatividad de 0.05.

El primer procedimiento es plantear las hipótesis, que podrían enunciarse como sigue:

H_0: La concentración media de monóxido de carbono en aire tomada de una muestra de datos, es igual a la reportada en los registros ?.

H_1: La concentración media de monóxido de carbono en aire tomada de una muestra de datos, es diferente a la reportada en los registros ?.

Se plantean las hipótesis estadísticas

$$H_{0:}\ \mu = \overline{X}$$

$$H_{1:}\ \mu \neq \overline{X}$$

Test a utilizar: $\boldsymbol{Z} = \dfrac{\overline{X} - \mu}{\sigma^2}$

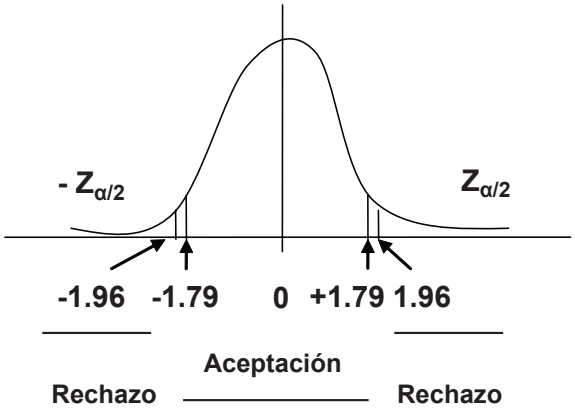
Aplicación del test: $\boldsymbol{Z} = \dfrac{10.43\ mg/\ m^3 - 10.00\ mg\ /\ m^3}{0.24\ mg\ /\ m^3} = 1.79\ mg/m^3$

Se busca 1.79 en la tabla de valores de ***Z*** a un nivel de significatividad de $\boldsymbol{\alpha} = 0.05$

Entonces como: $P(\overline{X} \geq 10.43\ mg\ /m^3) = P(\overline{Z}) \geq 1.79 = 0.0367$

El valor para una cola de la distribución corresponderá entonces 0.0367, como son dos colas en la ditribución $\boldsymbol{Z} = 0.0367+0.0367 = 0.0734$

Como ***Z(α)*** a un nivel de significatividad de 0.05 es 1.96, se compara este valor al valor obtenido anteriormente. Como ***P(Z(α) > Z***.



Conclusión: Como 0.0734 > 0.05 se acepta la $\boldsymbol{H_0}$, por lo que el promedio de la concentración de monóxido de carbono obtenido en la muestra es igual a la media de la población registrada.

Ejemplo 20. Se compara una muestra de veinticinco datos de temperaturas corporales de cangrejos intersticiales, con respecto a la temperatura ambiente promedio (24.3ºC) a un nivel de error del 5%. Los datos son los siguientes:

25.8, 24.6, 26.1, 22.9, 25.1, 27.3, 24.0, 24.5, 23.9, 26.2, 24.3, 24.6, 23.3, 25.5, 28.1, 24.8, 23.5, 26.3, 25.4, 25.5, 23.9, 27.0, 24.8, 22.9, 25.4.

El primer procedimiento es plantear las hipótesis, que podrían enunciarse como sigue: la media es de 25.03°C

H_0: La temperatura corporal promedio de la muestra de camarón es igual a la temperatura ambiente promedio?.

H_1: La temperatura corporal promedio de la muestra de camarón es diferente a la temperatura ambiente promedio?.

Se plantean las hipótesis estadísticas

$H_0: \mu = 24.3$ °C

$H_1: \mu \neq 24.3$ °C

$n = 25$

$\overline{X} = 25.03$ °C

$S^2 = 1.80$ °C

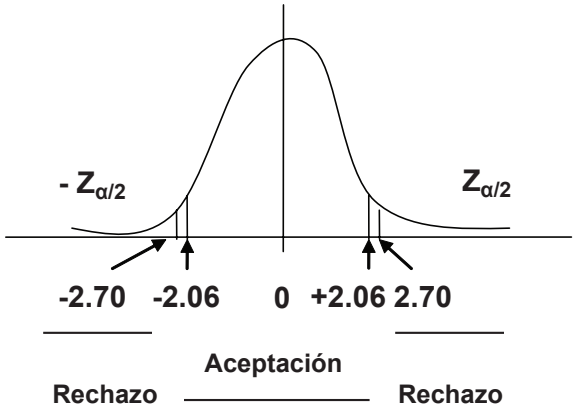
$S_{\bar{x}} = \sqrt{\frac{1.80\ °C}{25}} = 0.27°C$

$S_{\bar{x}} = \sqrt{\frac{S^2 x}{n}} = 0.27°C$ error estándard

Test a utilizar: $t = \frac{\overline{X} - \mu}{S_{\bar{x}}}$

Aplicación del test: $t = \frac{25.03°C - 24.3\ °C}{0.27\ °C} = 2.704°C$

Se busca en la tabla de valores de t el 2.704 a un nivel de significatividad de $\alpha = 0.05$ y a 24 gl (grados de libertad $n-1$ es decir 25-1). El valor obtenido es 2.064.



Conclusión: Como $t(\alpha)$ a un nivel de significatividad de 0.05 es 2.064 y este es menor que 2.704 calculado, se rechaza H_1 y se acepta H_1 por lo que la muestra de la temperatura media de los cangrejos que provienen de la zona intersticial es diferente a la del medio ambiente.

3.5.1.1 Prueba de hipótesis de una muestra para la varianza poblacional

Así como se puede comparar una media muestral a una media poblacional teórica, el proceso es idéntico para la varianza (figura 31).

$$H_0 : \sigma^2 = \text{valor hipotético}$$

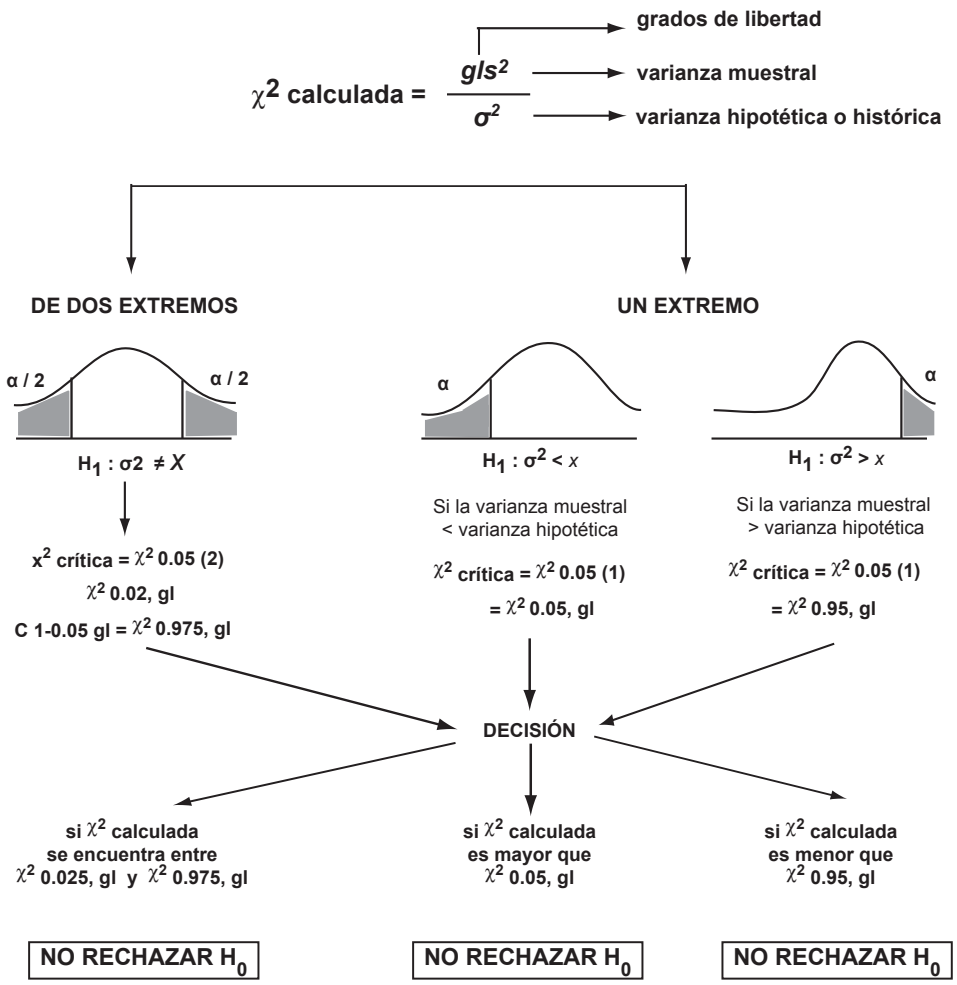


Figura 31. Prueba de hipotésis paramétrica para una muestra (varianza observada *vs* Varianza hipotética o histórica

Ejemplo 21. Se quiere saber si la varianza obtenida del tiempo de disolución de un alimento artificial en agua (2.6898 seg^2) es significativamente mayor a 1.5 seg^2. datos obtenidos de una muestra de 7 lecturas del tiempo de disolución del alimento artificial en agua.

H_0: La varianza obtenida del tiempo de disolución de un alimento artificial en agua es igual a 1.5 seg^2.

H_1: La varianza obtenida del tiempo de disolución de un alimento artificial en agua es mayor que 1.5 seg^2.

Ejemplo 21 (Continuación)

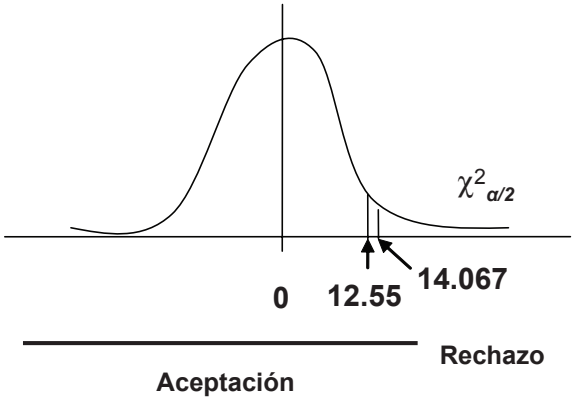
$H_0: s_x^2 = 1.5 \text{ seg}^2$

$H_1: s_x^2 > 1.5 \text{ seg}^2$

Test a utilizar: $\chi 2 = \dfrac{(gl)\,(s^2_x)}{\sigma^2_x}$ donde: gl son los grados de libertad

Aplicación del test: $\chi^2 = \dfrac{(7)\,(2.6898)}{1.5} = \dfrac{18.8288 \text{ sec}^2}{1.5 \text{ sec}^2} = 12.553$

Ahora se busca el valor de χ^2 de la tabla a un nivel de significatividad de 0.05 y para 7 grados de libertad que es 14.067



Conclusión: como 12.553 < 14.067, no se rechaza H_0 a un nivel de significatividad de 0.05 por lo que la varianza obtenida del tiempo de disolución en el alimento artificial en agua no es significativamente mayor a 1.5 seg^2.

3.5.2. Pruebas de Hipótesis con Dos Muestras

Dentro de los procedimientos estadísticos más comúnmente empelados está la comparación de dos muestras para inferir si existen diferencias entre dos poblaciones. Como en el caso de una sola muestra las pruebas de hipótesis para dos muestras pueden ser utilizadas para comparar dos medias, dos varianzas, dos proporciones, etc.

Al igual que para los casos de una sola muestra; las pruebas de hipótesis para dos muestras donde se hacen inferencias acerca de la población son denominados métodos paramétricos.

Por otra parte de acuerdo con la naturaleza de las muestras en términos de estrecha interdependencia podemos distinguir entre **muestras dependientes** y **muestras independientes**; las primeras son reconocidas en algunas instancias cuando cada observación de la primera muestra esta de alguna manera correlacionada con una o más observaciones de la segunda muestra, o en su caso, cuando los elementos de la segunda muestra son el resultado final de los elementos de la primera muestra

después de un tratamiento determinado. De manera contraria las muestras independientes únicamente comparten la variable, sin embargo estas podrían o no estar correlacionadas, o en su caso ser tomadas bajo condiciones distintas.

3.5.2.1. Pruebas paramétricas para muestras independientes

3.5.2.1.1. Comparación entre dos varianzas muestrales

Este análisis lo que busca a saber es si las varianzas y por supuesto sus desviaciones de dos muestras (test F) o varias muestras (test Barlett), no será tratado en este libro), son las estimaciones de una misma varianza σ^2 o de una misma σ. Para realizar este test se supone la normalidad de la distribución. Este test consiste a someter a una prueba de verdad la hipótesis principal de igualdad de varianzas (figura 32).

Si la hipótesis alternativa propuesta es $\sigma^2_1 < \sigma^2_2$, la resolución del test y las reglas de decisión son iguales, pero hay que cambiar el numerador y el denominador de la relación ***F*** y los números de ***gl*** del valor crítico de ***Fα*** de la primera muestra y los números de ***gl*** del valor crítico de ***Fα*** de la segunda muestra.

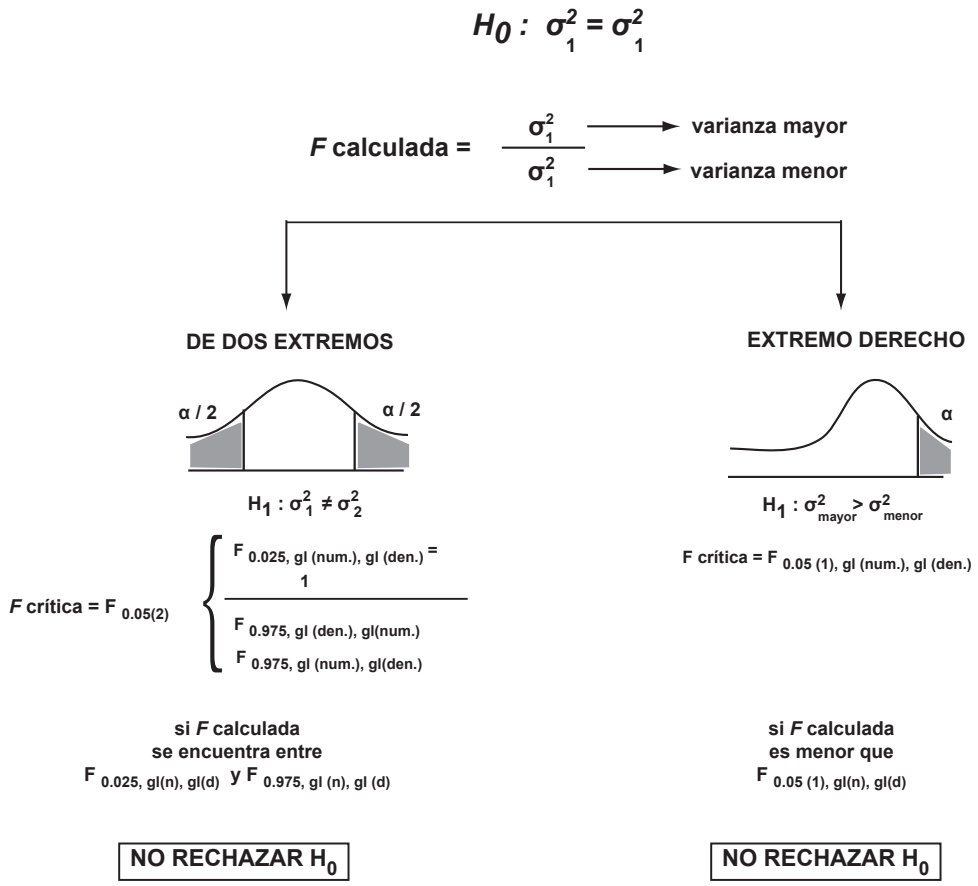


Figura 32. Prueba de hipótesis para dos muestras
(Varianza muestra 1 *vs* varianza muestra 2)

Así F calculada = $\frac{\sigma_1^2}{\sigma_2^2}$ *equivale a decir F calculada =* $\frac{S_{x1}^2}{S_{x2}^2}$ *de las muestras comparadas.*

Los grados de libertad también tendrán que ser cambiados de orden para poder ser leídos en las tablas de ***F***.

Ejemplo 22. En un estudio se hicieron determinaciones de amilasa en el plasma de una población de personas hospitalizadas, y otro de personas no hospitalizadas, la mayor varianza se presentó en las primeras personas; los resultados son los siguientes:

Hospitalizados	No hospitalizados
n_1= 22	n_2=15
S^2_1= 1600	S^2_2= 1225

Condiciones de aplicación: los datos provienen de muestras independientes, distribuidas normalmente.

H_0: La varianza en la determinación de amilasa en el plasma de una población de personas hospitalizadas es igual al de las personas no hospitalizadas.

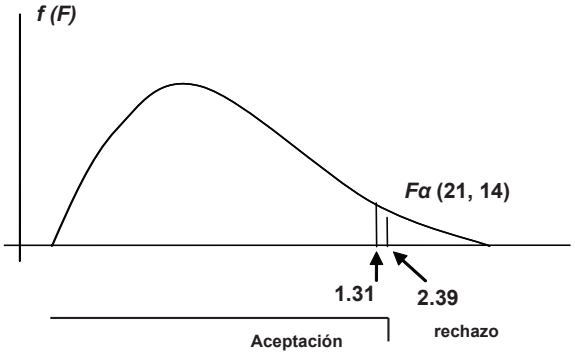
H_1: La varianza en la determinación de amilasa en el plasma de una población de personas hospitalizadas es mayor al de las personas no hospitalizadas.

$$H_0: \sigma^2_1 = \sigma^2_2$$

$$H_1: \sigma^2_1 > \sigma^2_2$$

Test a utilizar: $F = \frac{S^2_{x1}}{S^2_{x2}}$

Aplicación del Test: $F = \frac{1600}{1225} = 1.31$



Ahora se busca el valor de ***F*** de la tabla a un nivel de significatividad de 0.05 y para 21 (22-1) y 14 grados de libertad (15-1), el valor obtenido es 2.39, por lo tanto como ***F*** calculado es menor a ***Fα***(2.39) se acepta la H_0 , por lo que la determinación de la amilasa en el plasma de las personas hospitalizadas no es significativamente mayor que el de las personas no hospitalizadas.

3.5.2.1.2. Comparación entre dos medias muestrales independientes, bajo el supuesto de homogeneidad de varianzas (grandes muestras independientes n>30 y pequeñas muestras independientes n< 30)

Grandes muestras. Si $X_1, X_2,...X_n$ son variables aleatorias, independientes y normales, su suma o diferencia siguen una distribución normal, donde la esperanza matemática es la suma o la diferencia de las esperanzas matemáticas.

$$Z=X_1+X_2+...+X_n \quad o \quad Z=X_1-X_2-...-X_n$$

$$E(Z)=E(X_1)+E(X_2)+...+E(X_n)$$

$$E(Z)=E(X_1)-E(X_2)-...-E(X_n)$$

Por lo tanto la varianza es igual a la suma de las varianzas de las variables de origen.

$$\sigma^2=\sigma^2(X_1)+\sigma^2(X_2)+...+\sigma^2(X_n)$$

Si se considera entonces que: dos poblaciones de origen, tienen medias idénticas, se podrían comparar mediante la siguiente fórmula:

$$Zc = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{Sx_1^2}{n_1} + \frac{Sx_2^2}{n_2}}}$$

La H_0 se somete a una prueba de verdad después de poner formular H_1, la cual se puede formular de tres formas:

a) $H_1 : \mu_1 \neq \mu_2$ (test bilateral)

Si H_0 es verdad, el valor de Z calculado tiene una probabilidad $1-\alpha$ de situarse en el intervalo siguiente: $P(-Z_{/2} < Z_c < +Z_{/2}) = 1-\alpha$

Hay dos reglas de decisión:

- Si Z_c se encuentra entre los 2 valores críticos $-Z_{\alpha/2}$ y $+Z_{\alpha/2}$ H_0 es aceptada y se corre un riesgo β del cual no se conoce el nivel.
- Si Z_c se encuentra al exterior del intervalo mencionado, H_0 es rechazada y H_1 aceptada al nivel de significatividad α, del cual se conoce el nivel.

b) $H_1 : \mu_1 > \mu_2$ (test unilateral)

Si H_0 es verdad, el valor de Z calculado tiene una probabilidad $1-\alpha$ de situarse en el intervalo siguiente: $P(-Z > Z_c) = 1-\alpha$

Hay dos reglas de desición:

- Si Z_c es inferior al valor crítico $Z_{\alpha,}$ H_0 es aceptada y se corre un riesgo del cual no se conoce el nivel.
- Si Z_c es superior al valor crítico Z_{α}, H_0 es rechazada y H_1 aceptada y se corre un riesgo de error β.

c) $H_1 : \mu_1 < \mu_2$ (test unilateral)

Si H_0 es verdad, el valor de Z calculado tiene una probabilidad $1-\alpha$ de situarse en el intervalo siguiente: $P(-Z_\alpha > Z_c) = 1 - \alpha$

Hay dos reglas de decisión:

- Si Z_c es superior al valor crítico $-Z_\alpha$, H_0 es aceptada y se corre un riesgo β del cual no se conoce el nivel.
- Si Z_c es inferior al valor crítico $-Z_\alpha$, H_0 es rechazada y H_1 aceptada y se corre un riesgo de error β.

La figura 33, muestra la prueba de hipótesis paramétrica para dos muestras independientes, suponiendo homogeneidad de varianzas, se muestra el test a utilizar para grandes muestras y pequeñas muestras.

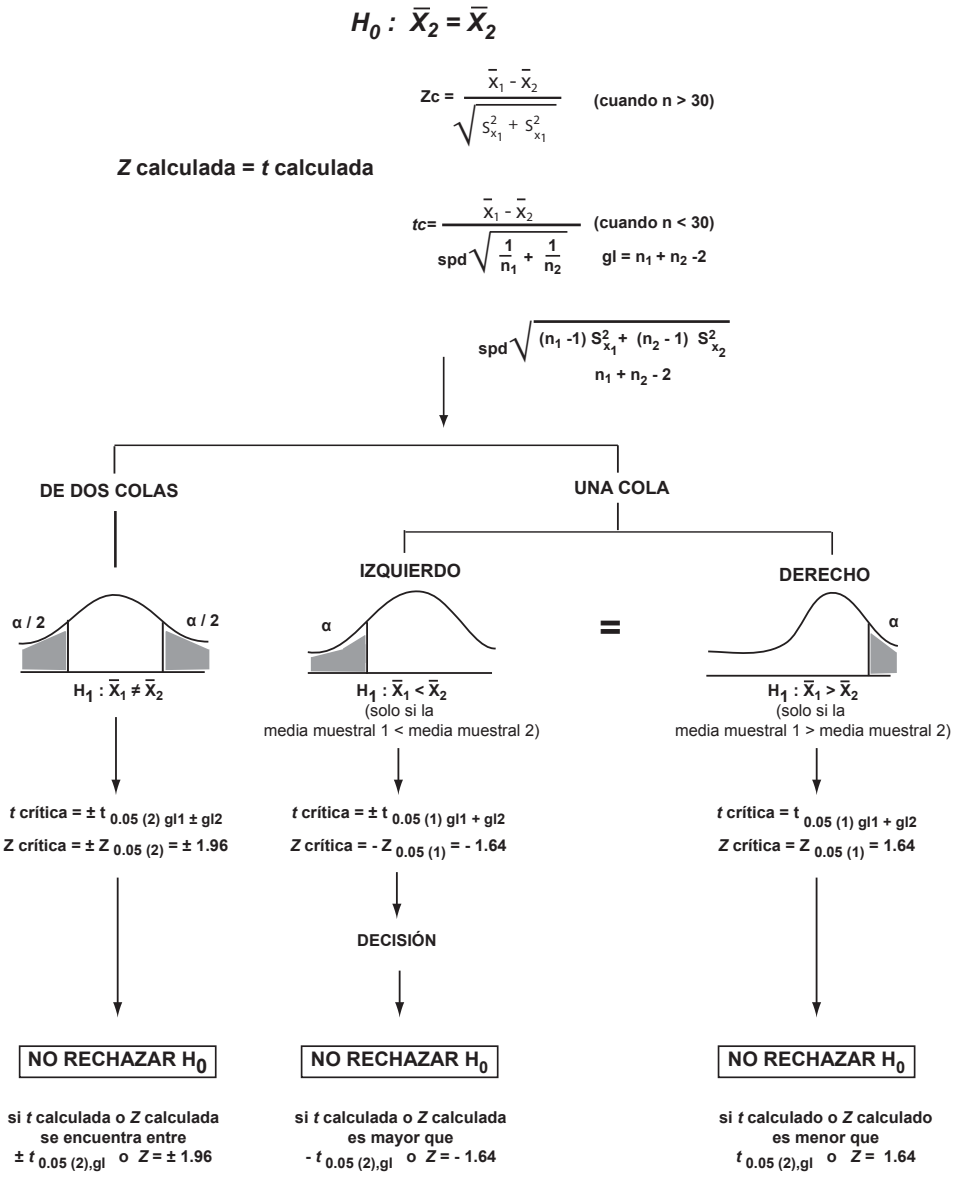


Figura 33. Prueba de hipótesis paramétrica para dos muetras independientes, suponiendo igualdad de varianzas (media muestra 1 *vs* media muestra 2)

Ejemplo 23. Se estudia el dimorfismo sexual del pájaro (*Bonasa umbrella*) a través de muestras obtenidas por caza. Los datos fueron obtenidos al azar en Canadá. Se obtienen dos muestras, la primera compuesta por machos juveniles y la segunda por hembras juveniles, se tiene como variable la longitud de la cola como medida para definir el dimorfismo sexual. Los resultados fueron los siguientes:

Hembras	Machos
$n_1 = 50$	$n_2 = 67$
$\bar{X}_1 = 158.86$ mm	$\bar{X}_2 = 134.46$ mm
$S^2x_1 = 37.18$ mm	$S^2x_2 = 25.92$ mm
$S_{x1} = 6.09$ mm	$Sx_1 = 5.09$ mm

La pregunta impuesta en este ejemplo es saber si las medias de estas 2 muestras tomadas independientemente difieren una de otra de manera altamente significativa ($\alpha < 0.01$)?.

Condiciones de aplicación: los datos provienen de muestras independientes, distribuidas normalmente.

Planteamiento de hipótesis:

H_0: La media de la longitud de la cola de las aves machos es igual a la de las aves hembras.

H_1: La media de la longitud de la cola de las aves machos es diferente a la de las aves hembras.

H_0: $\mu_1 = \mu_2$

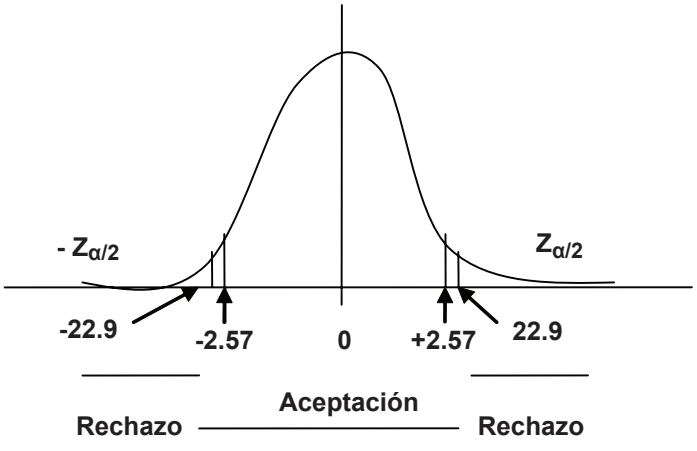
H_1: $\mu_1 \neq \mu_2$

Test estadístico:

$$Zc = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{Sx_1^2}{N_1} + \frac{Sx_2^2}{N_2}}}$$

$$Zc = \frac{158.86 \text{ mm} - 134.66 \text{ mm}}{\sqrt{\frac{37.18 \text{ mm}}{50} + \frac{25.92 \text{ mm}}{67}}} = 22.9$$

Decisión estadística: como ***Zc*** es superior a 2.57, que es el valor de ***α*** a 0.01, se rechaza H_0 y se acepta H_1.



Interpretación: La longitud de de la cola de los pájaros machos juveniles y hembras de *Bonasa umbrellus* es diferente del de las hembras juveniles; por lo tanto hay dimorfismo sexual en esta ave.

Pequeñas muestras. En este caso, se utiliza el test de ***t***, en donde:

$$tc = \frac{\bar{X}_1 - \bar{X}_2}{Spd\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad para\ gl = n_1 + n_2 - 2$$

S_{pd} corresponde a la media ponderada de las varianzas de 2 muestras, esta media es una estimación muy precisa de la varianza σ^2 y se expresa:

$$S^2_{pd} = \frac{(n_1 - 1)\, S^2 x_1 + (n_2 - 1) S^2 x_2}{n_1 + n_2 - 2}$$

Para obtener S_{pd} se tendría que sacar la raíz cuadrada del resultado. Para aplicar este test, las 2 poblaciones deben tener la misma varianza, si esta condición no se cumple se puede aplicar el test siempre y cuando los valores de n_1 y n_2 no sean muy diferentes. Para definir la aceptación o rechazo de la H_0 consultar la tabla respectiva. También se puede utilizar la ecuación descrita en la siguiente figura, en cualquiera de los casos el resultado es el mismo.

Ejemplo 24. Se quiere probar si existen diferencias significativas en los tiempos (minutos) de saturación en sangre de dos tipos de medicamentos suministrados a dos grupos de organismos (Un medicamento por grupo).

Grupo A Tiempo (min)	Grupo B Tiempo (min)
8.8	9.9
8.4	9.0
7.9	11.1
8.7	9.6
9.1	8.7
9.6	10.4
	9.5

Condiciones de aplicación: los datos provienen de muestras independientes, con datos < a 30.

Planteamiento de hipótesis:

H_0: Los tiempos de saturación en sangre de dos tipos de medicamentos suministrados en el grupo A y B son iguales.

H_1: Los tiempos de saturación en sangre de dos tipos de medicamentos suministrados en el grupo A y B son diferentes.

H_0: $\mu_1 = \mu_2$

H_1: $\mu_1 \neq \mu_2$

Test estadístico: $tc = \frac{\bar{X}_1 - \bar{X}_2}{Spd\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ para $gl = n_1 + n_2 - 2$

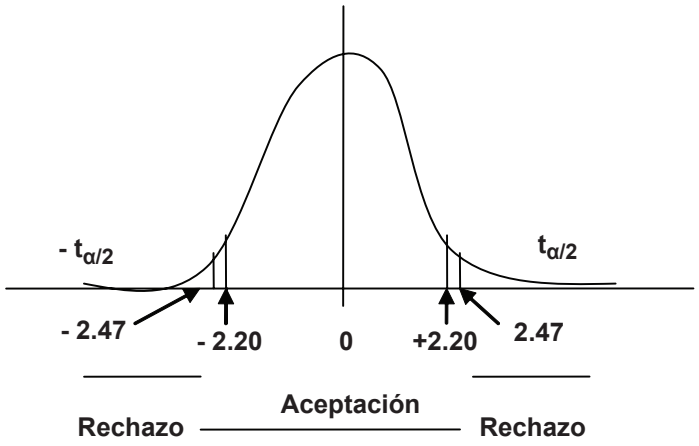
$$s^2_{pd} = \frac{(n_1 - 1)\, S^2 x_1 + (n_2 - 1) S^2 x_2}{n_1 + n_2 - 2}$$

Calculo del test: $S^2_{pd} = \frac{(8.75 \text{ min} - 9.74 \text{ min})}{6 + 7 - 2} = 0.5193$

Ejemplo 24 (Continuación)

Para poder calcular el test de ***t***, ahora hay que obtener S_{pd}, *por lo que se hace necesario calcular la raíz cuadrada de* S^2_{pd}.

$$tc = \frac{(8.75 \text{ min} - 9.74 \text{ min})}{0.7206\sqrt{\frac{1}{6}+\frac{1}{7}}} = 2.475$$



Decisión estadística: como ***tc*** es superior a $t_{\alpha/2}$ para 11 grados de libertad (2.201), se rechaza H_0 y se acepta H_1, por tanto los tiempos de saturación en sangre de dos tipos de medicamentos suministrados en el grupo A y B son diferentes.

3.5.2.1.3 *Comparación entre dos medias muestrales independiente, bajo el supuesto de varianzas diferentes (pequeñas muestras independientes n<30).*

En este caso se utiliza el test de ***t***, llamado test crítico modificado (figura 34), en donde:

$$tcm = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_{x_1}^2}{n_1}+\frac{S_{x_2}^2}{n_2}}}$$

En este caso, los grados de libertad se calculan de la siguiente manera:

$$gl = \frac{\left[\left(\frac{S^2 x_1}{n_1}\right)+\left(\frac{S^2 x_2}{n_2}\right)\right]^2}{\frac{\left(\frac{S^2 x_1}{n_1}\right)^2}{n_1 - 1}+\frac{\left(\frac{S^2 x_2}{n_2}\right)^2}{n_2 - 1}}$$

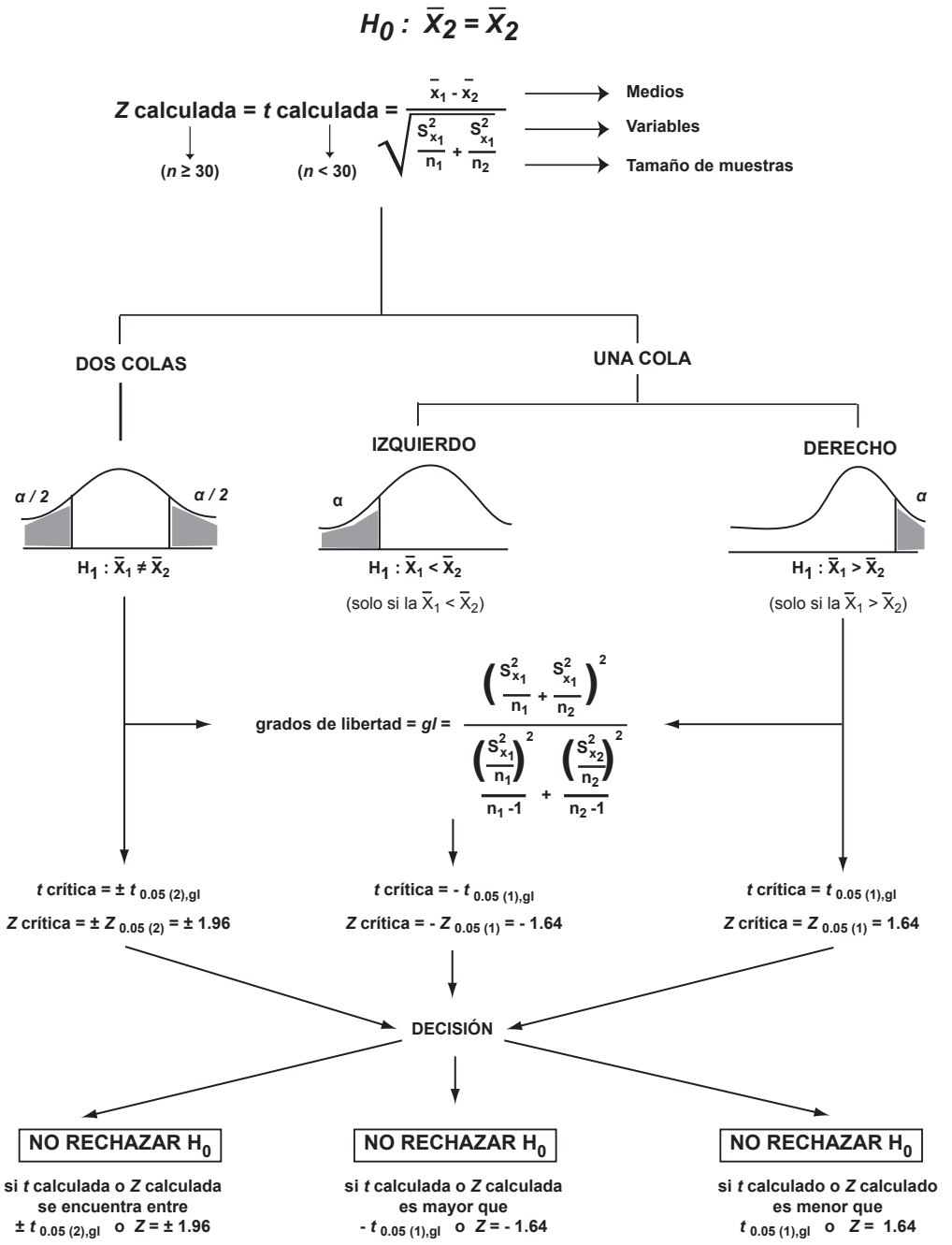


Figura 34. Prueba de hipótesis paramétrica para dos pequeñas muestras independientes, suponiendo desigualdad de varianzas (media muestra 1 *vs* media muestra 2).

Ejemplo 25. En un estudio inmunológico, Elie y Lamoreaux (1974) desgarraron la piel de dos lotes de ratones. El primer lote de 10 ratones no recibió ningún tratamiento. El segundo lote de 10 ratones, ha sido tratado por un antígeno de transplante incubado en suero normal. La sobrevivencia de ratones fue medida en días y los resultados de esta experiencia son los siguientes:

Testigos	Tratados
n_1=10	N_2=10
$\bar{X}_1$=9.42 días	$\bar{X}_2$=13.36 días
S_{X_1}=0.49	S_{X_1}=1.63

Se plantea la pregunta de saber si la sobrevivencia de los ratones se prolonga con el tratamiento.

Condiciones de aplicación: los datos provienen de muestras independientes, con datos < a 30, varianza diferente por lo que siguen la ley de ***t-student***.

Planteamiento de hipótesis:

H_0: La sobrevivencia promedio de los ratones tratados es igual a la de los ratones no tratados.

H_1: La sobrevivencia promedio de los ratones tratados es mayor que la de los ratones no tratados

$H_0: \bar{x}_1 = \bar{x}_2$

$H_1: \bar{x}_1 < \bar{x}_2$

Test estadístico:

$$tcm = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S^2x_1}{n_1} + \frac{S^2x_2}{n_2}}}$$

Calculo de grados de libertad:

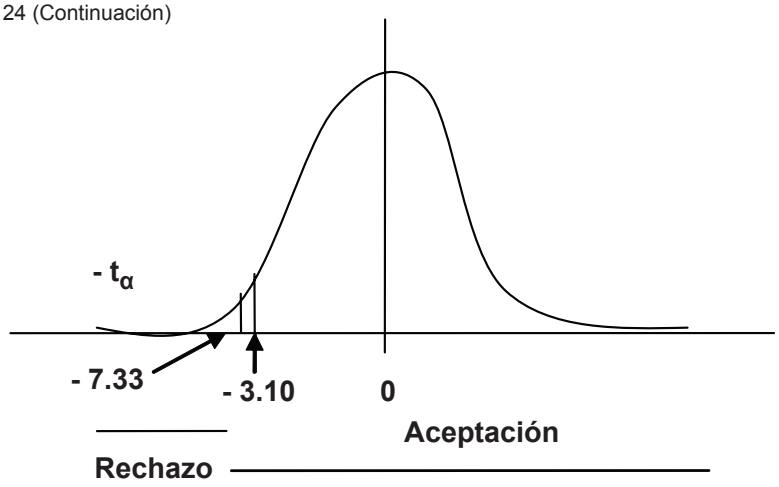
$$gl = \frac{\left[\left(\frac{S^2x_1}{n_1}\right) + \left(\frac{S^2x_2}{n_2}\right)\right]^2}{\frac{\left(\frac{S^2x_1}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S^2x_2}{n_2}\right)^2}{n_2 - 1}}$$

Cálculo del test:

$$tcm = \frac{9.42 - 13.36}{\sqrt{\frac{0.24}{10} + \frac{2.65}{10}}}$$

$$gl = \frac{\left[\left(\frac{0.24}{10}\right) + \left(\frac{2.65}{10}\right)\right]^2}{\frac{\left(\frac{0.24}{10}\right)^2}{9} + \frac{\left(\frac{2.65}{10}\right)^2}{9}} = 10.66$$

Ejemplo 24 (Continuación)



Decisión estadística: como ***tcm*** es inferior al valor crítico de t_α para 11 grados de libertad (-7.33<-3.10), se rechaza H_0 a un nivel de significatividad de 0.05, por lo que la sobrevivencia de ratones es prolongada con el tratamiento a los antígenos que fueron encubados en un suero normal.

3.5.2.4. Comparación de medias de 2 muestras apareadas

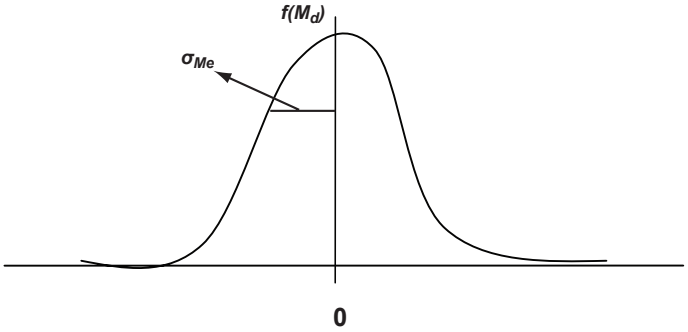
Se basa en el análisis de las diferencias observadas a nivel de cada par de elementos: $di=x_{i1}-x_{i2}$. La muestra de n diferencias d_i, tiene una media $\bar{d}$ y una varianza de S_d^2. Si n es suficientemente grande, la variable M_d, es igual a $M_1 - M_2$ y corresponde a los valores tomados por $\bar{d}$ a nivel de diferentes muestras posibles de talla n, obedecen a una ley normal de media $\mu_{Md} = M_d = E(M_1)-E(M_2) = \mu_1-\mu_2$;

y la desviación estándar $\sigma_{Md} = \frac{\sigma_d}{\sqrt{n}}$

Por lo tanto la variable centrada obedecerá también a una distribución normal.

$Z_{md} = \frac{M_d - \mu_d}{\sigma_{md}}$; sin embargo, se utiliza normalmente la variable t_{Md},que obedece a la ley de ***t-Student*** con n= 1 gl (figura 35).

$t_{md} = \frac{M_d - \mu_d}{S_{\bar{d}}}$



Se somete entonces la comparación de medias a una prueba de hipótesis de igualdad de medias. H_0: $\mu_1=\mu_2$ o $\mu_d= \mu_1 - \mu_2=0$; la variable t_{md} es:

$$t_{md} = \frac{M_d}{S_{\bar{d}}}$$

y para 2 muestras apareadas es: $t_{\bar{d}} = \frac{\bar{d}}{S_{\bar{d}}}$; $\bar{d}$ = diferencia media.

$$\text{Error tipo} = S_{\bar{d}} = \frac{S_d}{\sqrt{n}}; \; S_d = \sqrt{\frac{\sum^{n} (d_i - \bar{d})^2}{n-1}}$$

H_1: $\mu_1 \neq \mu_2$ *Prueba bilateral: la H_0 se rechaza cuando el valor calculado: ($|t_{\bar{d}}| \geq t_{\alpha/2}$)*

H_1: $\mu_1 > \mu_2$ *Prueba bilateral: la H_0 se acepta cuando el valor calculado: ($t_{\bar{d}} < +t_{\alpha}$)*

H_1: $\mu_1 < \mu_2$ *Prueba bilateral: la H_0 se acepta cuando el valor calculado: ($t_d > -t_{\alpha}$)* ***gl =n-1***

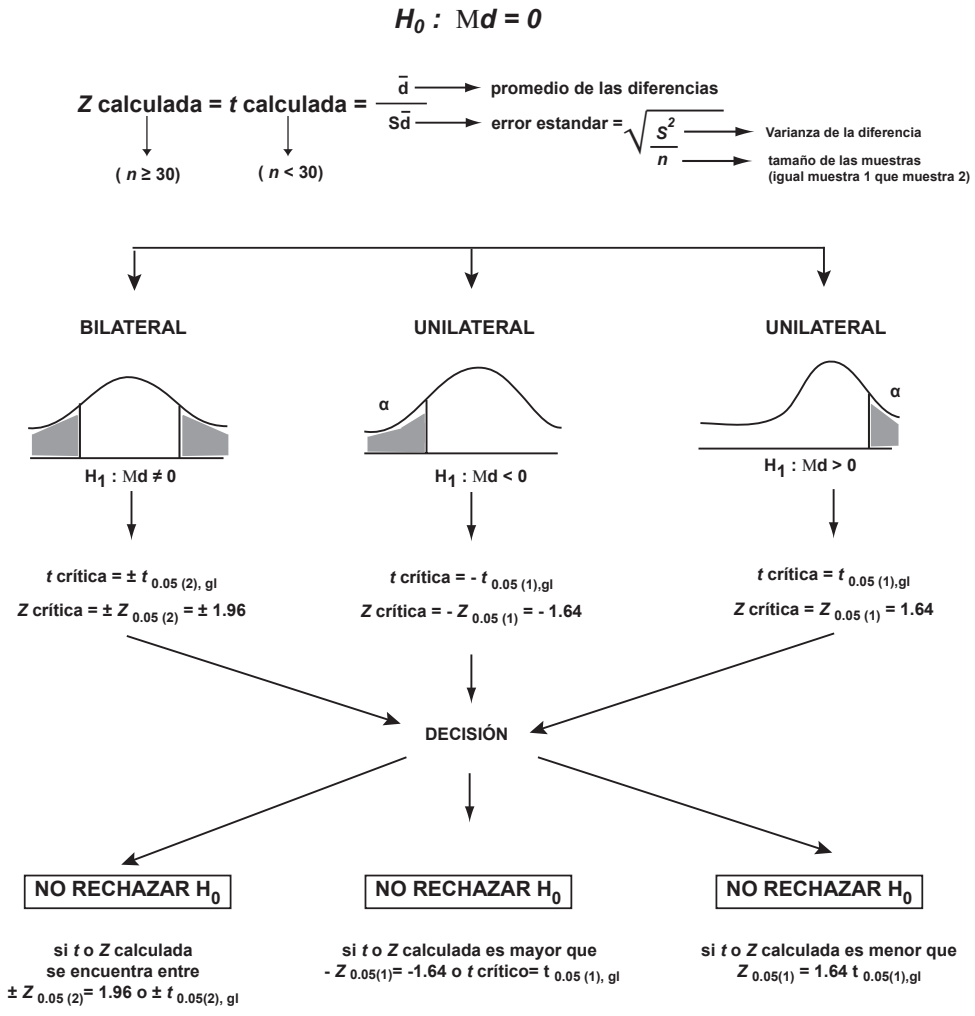


Figura 35. Prueba de hipótesis paramétrica para dos muestras dependientes

Ejemplo 26. En un estudio morfométrico se quiso probar si existían diferencias significativas entre las longitudes de las extremidades delanteras con respecto a las traseras en venados, midiéndose 10 organismos y considerando un 5% de error.

Conejos	Longitud patas delanteras (cm) d_1	Longitud patas traseras (cm) d_2	Diferencia $(d_1 - d_2)$	$(di - \bar{d})^2$
1	142	138	4	0.49
2	140	136	4	0.49
3	144	147	-3	39.69
4	144	139	5	2.89
5	142	143	-1	18.49
6	146	141	5	2.89
7	149	143	6	7.29
8	150	145	5	2.89
9	142	136	6	7.29
10	148	146	2	1.69

Los resultados de las medidas se presentan en la tabla:

H_0: La longitud de las extremidades delanteras y traseras no presentan diferencia

H_1: La longitud de las extremidades delanteras con respecto a las traseras es diferente

H_0: $\mu d_1 = \mu d_2$

H_1: $\mu d_1 \neq \mu d_2$

Condiciones de aplicación. La información proviene de la misma muestra de venados por lo que son muestras apareadas o dependientes, se aplica un test de tμd.

Test a aplicar: $t_d = \frac{\bar{d}}{S_{\bar{d}}}$ donde es la diferencia media entre d_1 y d_2.

Aplicación del test: $\bar{d} = 3.3$ cm

$Sd^2 = 9.34444$ cm^2

$S\bar{d} = 0.97$ cm

$$t = \frac{\bar{d}}{Sd} = \frac{3.3}{0.97} = 3.402$$

Decisión estadística: como ***t*** a un nivel 0.05 y 9 grados de libertad (10-1) es igual a 2.262 y es inferior a 3.402, se rechaza la H_0, y se acepta H_1 por lo que existe una diferencia significativa entre la longitud de las patas delanteras y traseras de la muestra de 10 venados.

4 Pruebas de Normalidad

Literatura sugerida:

Sokal R. y F. J. Rholf. 2000. Biometry (Tercera edición) Ed. Freeman. (Pag. 118-125, 697-708)

Zar. J. H., 1999. Bostatistical Analysis (4 edición) Prentice Hall. Estados Unidos. 663 p. (Pag. 462-469).

Scherrer B. 1984. Boestatistique. Gaetan Morin editeur. (Pag. 466-479)

En algunas ocasiones es necesario probar la hipótesis de que una muestra viene de una población que sigue una distribución normal. La importancia de saber si una muestra viene de una población distribuida de acuerdo a una distribución normal, radica básicamente en que algunos métodos de inferencia paramétricos (basadas en el análisis de los parámetros de la población) como por ejemplo las pruebas ***t-Student*** y de análisis de varianza suponen que las muestras provienen de una distribución normal, por lo tanto si no se demuestra la "normalidad" de una muestra, es necesario emplear métodos alternativos no paramétricos que no se basen en el supuesto antes mencionado; y de esta forma evitar incurrir en errores de inferencia.

4.1 Prueba de Normalidad *Q-Q* (Cuantil-Cuantil)

La prueba de normalidad cuantil-cuantil, es un método gráfico que se basa en la distribución de frecuencias (para datos agrupados) donde la frecuencia acumulada representa las proporciones de la distribución normal, las cuales son transformadas a una escala de desviaciones estándar (***z***), denominadas equivalentes de la distribución normal (EDN's).

Para el caso de los datos no agrupados, estos se ordenan de forma ascendente, en donde el orden numérico asignado es transformado a proporciones de la distribución normal, seguidamente a los EDN's.

Con estos EDN's se elabora una gráfica de dispersión donde los limites superiores de cada marca de clase o los datos no agrupados, se disponen a lo largo del eje de las abcisas y los EDN's, a lo largo del eje de las ordenadas.

Como se observa en la figura 36 basada en la tabla de distribución de frecuencias del ejercicio 1, los puntos deben ajustarse a una línea recta, de no ocurrir lo anterior, podríamos decir que nuestros datos no se ajustan a la distribución normal.

Dichos gráficos de dispersión se conocen como gráficos cuantil-cuantil o ***Q-Q***, o como gráficos de cuantiles normales. Los gráficos ***Q-Q***, dan mejores resultados para muestras mayores a 50 datos. En algunos casos, cuando se tiene muestras menores a 50 una diferencia importante en un solo dato con respecto a los demás, generaría una desviación sustancial con respecto a la línea recta.

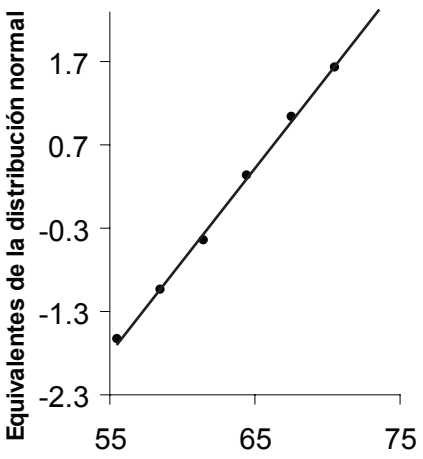


Figura 36. Dispersión de los limites superiores de cada intervalo de clase con los equivalentes de la distribución normal (EDN's) o gráfico cuantil-cuantil.

4.2. D'Agostino-Pearson K^2

En una revisión completa de algunos métodos disponibles D'Agostino (1986) concluyó que el procedimiento más aceptable para probar la hipótesis acerca de la normalidad de una seríe de datos es el descrito por D'Agostino y Pearson (1973). La hipótesis nula de la normalidad de la población es probada a partir del siguiente cálculo:

$$K^2 = Z^2_{g1} + Z^2_{g2}$$

Donde: Z^2_{g1} y Z^2_{g2} son estadísticos correspondientes a la simetría (*g1*) y curtosis (*g2*) respectivamente. La significancia de K^2 es determinada mediante su aproximación a la distribución chi-cuadrada, χ^2, con dos grados de libertad (En el caso particular de esta prueba lo grados de libertad se mantendrán constantes. De acuerdo con D'Agostino esta prueba trabaja bien con muestras con $n \geq 20$.

La hipótesis nula H_0: La población muestreada se encuentra normalmente distribuida; H_1: La población muestreada no se encuentra normalmente distribuida) es rechazada si el estadístico calculado, K^2, es mayor al estadístico crítico, $\chi^2 = 5.99$ ($\alpha = 0.05$, $gl = 2$); lo cual sugiere asimetría, lepto o platicurtosis o todas al mismo tiempo; e indica que la muestra no viene de una distribución normal.

Para ver el cálculo preciso de esta prueba, un ejemplo se detalla en Zar (1999) para quienes quieren conocer de una manera más amplia su utilización.

El procedimiento puede ser descrito a través de cuatro pasos principales: (1) Cálculo de la simetría, g_1, y curtosis, g_2; (2) "Normalización" de la simetría, Zg_1; (3) "Normalización" de la curtosis, Zg_2; (4) Prueba de hipótesis para probar normalidad, estimación del estadístico muestral, K^2, aproximado a la distribución normal.

Es importante resaltar que este método ha caído en desuso y en esta obra sólo se muestra como debe de ser utilizado en caso de que el usuario tenga duda sobre los datos que analiza. Su subutilización se

debe a tres causas: la primera es que resulta bastante complicado y consume mucho tiempo, la segunda es que se han encontrado métodos gráficos más fáciles que dan una idea aproximada de la forma de la distribución de frecuencias, y la tercera es que en los últimos años se ha dado menos énfasis a la forma de las distribuciones y se ha utilizado el supuesto de que "grandes muestras, mayores a 30 datos *n* mayor a 30" siguen una distribución normal, lo que no sucede con muestras menores a 30 datos.

4.3. Pruebas de Bondad de Ajuste

Esta prueba se utiliza para determinar si una muestra de valores observados, de alguna manera aleatoria es compatible o no con la hipótesis de que se extrajo de una población de valores que está distribuida normalmente. Consiste en colocar los valores en categorías o intervalos de clase y observar la frecuencia de ocurrencia de los valores en cada categoría. Entonces se aplica la prueba de distribución normal con el fin de determinar las frecuencias que podrían esperarse para cada categoría, si la muestra hubiera provenido de una distribución normal.

Si la diferencia entre lo que se observó y lo que se espera (dado que el muestreo fue de una distribución normal) es demasiado grande como para ser atribuida al azar, se concluye que la muestra no provino de una distribución normal. Si la diferencia es de tal magnitud que pudo haber intervenido el azar, se concluye que la muestra pudo haber provenido de una distribución normal.

Las pruebas de bondad de ajuste se basan en la distribución χ^2 propuesta por Karl Pearson (1935), con base en una muestra de tamaño *n* (comúnmente variables cualitativas) de un muestreo aleatorio, los pasos a seguir son los siguientes:

Planteamiento de las hipótesis

H_0: La distribución de frecuencias entre categorías es homogénea.

H_1: La distribución de frecuencias entre categorías es heterogénea.

	Categoría $_{1\,(k)}$	**Categoría $_{2\,(k)}$**	**...Categoría $_{k}$**
Frecuencia observada	O_1	O_2	$...O_i$
Frecuencia esperada	$(O_1+O_2)/k=E_1$	$(O_1+O_2)/k=E_2$	$...(O_1+O_2)/k=E_i$

*k = número de categorías

Establecer estadístico teórico (valor crítico)

$$\chi^2_{\alpha,v} = \chi^2_{0.05,\ k\ -1=2-1=1}$$

Establecer estadístico muestral (valor calculado)

$$\chi^2_{calculado} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

Conclusión

Si el estadístico muestral es mayor que el estadístico teórico rechazar la hipótesis nula.

Ejemplo 27. Supóngase que un investigador que realiza un estudio de hospitales en los Estados Unidos, reúne datos sobre una muestra de 250 hospitales la cual le permitió calcular la tasa de ocupación (proporción) de los pacientes internados. Los resultados son los siguientes:

Tasa de ocupación	No. de hospitales
< 40	16
40.0 a 49.9	18
50.0 a 59.9	22
60.0 a 69.9	51
70.0 a 79.9	62
80.0 a 89.9	55
90.0 a 99.9	22
100 >	4

Pregunta: Proporcionan estos datos evidencia suficiente que indique que la muestra no provino de una población normalmente distribuida?.

Planteamiento de las hipótesis.

H_0: Las datos se extrajeron de una población normalmente distribuida

H_1: Los datos no provienen de una población normalmente distribuida

Como el supuesto es la normalidad se deben conocer la media, varianza y desviación estándar de la muestra. Para este ejercicio, la media = 70.1, la varianza = 325.372 y la desviación estándar = 18.0381. Posteriormente se procede a obtener el valor de *Z* de cada intervalo (utilizando el limite inferior):

$$Z=\frac{40.0 - 70.1}{18.0381} = 1.67, \quad Z=\frac{50.0 - 70.1}{18.0381} = 1.11, \text{ etc}$$

Tasa de ocupación (Intervalos de clase)	$Z=\frac{\bar{X} - \mu}{\sigma_{\bar{x}}}$	Frecuencia relativa esperada	Proporción de la curva normal, valores de *Z*	Cálculo de frecuencias relativas esperadas	No. de hospitales. Frecuencia obervada	No. de hospitales. Frecuencia esperada
< 40		0.475		0.475	16	(250*4.75)/99.2=11.97
40.0 a 49.9	-1.67	0.0860	0.0475	(0.0475-0.1335)=0.056	18	(250*8.60)/99.2=21.67
50.0 a 59.9	-1.11	0.1542	0.1335	(0.1335-0.2877)=0.1542	22	(250*15.42)/99.2=38.86
60.0 a 69.9	-0.56	0.2083	0.2877	(0.2877-0.4960)=0.2083	51	(250*20.83)/99.2=52.49
70.0 a 79.9	-0.01	0.2048	0.4960	(0.4960-0.2912)=0.2048	62	(250*20.48)/99.2=51.61
80.0 a 89.9	0.55	0.1555	0.2912	(0.2912-0.1357)=0.1555	55	(250*15.55)/99.2=38.19
90.0 a 99.9	1.10	0.0872	0.1357	(0.1357-0.0485)=0.0872	22	(250*8.72)/99.2=21.98
100 >	1.66	0.0485	0.0485	0.0485	4	(250*4.85)99.2=12.22
Total		0.992			250	

La frecuencia relativa esperada se obtiene buscando el valor obtenido de *Z* en las tablas de áreas bajo la curva de distribución normal (*Z*) y el valor para cada intervalo de clase, corresponde a la diferencia de el valor de *Z* del último intervalo y el próximo inmediato. Finalmente se calculan las frecuencias esperadas para cada intervalo y se aplica el test.

$$\chi^2_{calculado} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

$$\chi^2_{calculado} = \frac{(16-11.97)^2}{11.97} + \frac{(18-21.67)^2}{21.67} + \frac{(22-38.86)^2}{38.86} + \frac{(51-52.49)^2}{52.49} + \frac{(62-51.61)^2}{51.61} + \frac{(55-38.19)^2}{38.19} + \frac{(22-21.98)^2}{21.98} + \frac{(4-12.22)^2}{12.22} = 23.34$$

Como $\chi^2_{(0.05,1)} = 6.63$ es menor que 23.34, se rechaza la hipótesis nula y se admite entonces que los datos no se distribuyen normalmente.

Ejemplo 28. En una investigación se planteó que existe la misma probabilidad de tener el mismo número de elementos por recuento de fitoplancton, según un 5% de error. Se tomaron 4 alícuotas de 1 ml, con un total de 100 células. Encontrándose la siguiente distribución de frecuencias:

	$Alícuota_1$	$Alícuota_2$	$Alícuota_3$	$Alícuota_4$	Total
Frecuencia observada	36	40	18	6	100
Frecuencia esperada	25	25	25	25	100

H_0: La distribución de frecuencias entre categorías es homogénea.

H_1: La distribución de frecuencias entre categorías es heterogénea.

El estadístico teórico es igual a $\chi^2_{0.05,\, k-1=4-1=3} = 7.81$

y el estadístico muestral:

$$\chi^2_{calculado} = \frac{(36-25)^2}{25} + \frac{(40-25)^2}{25} + \frac{(18-25)^2}{25} + \frac{(6-25)^2}{25} = 30.24$$

Por lo tanto, como el valor calculado es mayor que el valor crítico, se rechaza la hipótesis nula. Por lo que la probabilidad de tener el mismo número de elementos por recuento del fitoplancton no es aceptada y se admite que el número de elementos es diferente.

4.3.1. Tablas de Contingencia

En algunos casos los datos se colectan simultáneamente para dos variables, y por lo tanto podría ser deseable probar la hipótesis de que las frecuencias de ocurrencia en varias categorías de una variable son independientes de la segunda variable. El arreglo en el que pueden estar dispuestos los datos se denomina tabla de contingencia.

Un análisis estadístico a través de tablas de contingencia se basa en datos experimentales donde se tienen dos factores interrelacionados. A la tabla de contingencia también se le conoce como método de bloques y su procedimiento emplea la distribución de χ^2. La tabla de contingencia esta conformada por un número de renglones, *R* correspondientes al factor 1, y por un número de columnas, *C*, correspondientes al factor 2.

La hipótesis nula en una tabla de contingencia es probar que las variaciones de las frecuencias observadas en las filas son independientes de las frecuencias observadas en la columnas. De esta manera el procedimiento a seguir es el siguiente:

Planteamiento de las hipótesis.

H_0: La distribución de las frecuencias entre niveles es independiente de los factores.

H_1: La distribución de las frecuencias entre niveles no es independiente de los factores.

Establecer estadístico teórico (valor crítico).

$$\chi^2_{a,gl} = \chi^2_{0.05,\,(k-1)(f-1)}$$

	Factor$_1$	Factor$_2$	...Factor$_k$	∑Factores
Nivel$_{,1}$	$O_{1,1}$	$O_{2,1}$	...$O_{i,1}$	$\sum Ok_{i,1}$
Nivel$_{,2}$	$O_{1,2}$	$O_{2,2}$	... $O_{i,2}$	$\sum Ok_{i,2}$
Nivel$_{,f}$	$O_{1,j}$	$O_{2,j}$	... $O_{i,j}$	$\sum Ok_{i,j}$
∑Nivel	$\sum O_{1,i}$	$\sum Of_{2,i}$	$\sum Of_{i,j}$	∑TOTAL=N

Establecer estadístico muestral (valor calculado).

$$E_{i,j} = \frac{\sum Ok_{i,j} \left(\sum Of_{i,j}\right)}{N} \qquad \chi^2_{calculado} = \sum_{j=1}^{f} \sum_{i=1}^{k} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$$

Conclusión.

Si el estadístico muestral es mayor que el estadístico teórico rechazar la hipótesis nula. Por tanto la distribución de frecuencias entre los niveles no es dependiente de los factores.

Ejemplo 29. En la Isla de Grues en el estuario de Saint-Laurent, dos muestras de cazadores que se dedicaban a la captura de gansos blancos fueron obtenidas de manera independiente. La primera muestra estaba constituida con cazadores que pertenecían a un club de caza y la segunda de cazadores al servicio de proveedores de caza y pesca. El perfil de los dos grupos de cazadores es considerablemente diferente, sin embargo las condiciones de caza son las mismas. En 1974, 57.14 % de 35 cazadores de primer grupo y 68.67% de 83 cazadores del segundo grupo regresaron con las manos vacías (cero caza).

Tabla de contingencia			
	Cazadores proveedores	**Cazadores club**	**Total**
Cazadores con éxito	26	16	41
cazadore con las manos vacias	57(68.67%)	20(57.14%)	77
Total	83	35	118

Pregunta: el porcentaje de cazadores que regresaron con las manos vacías del primer grupo es significativamente diferente del segundo grupo?.

H_0: Las proporciones de los cazadores de ambos grupos son idénticas

H_1: : Las proporciones de los cazadores de ambos grupos son diferentes

$H_0: P_1 = P_2$

$H_1: P_1 \neq P_2$

Establecer estadístico teórico (valor crítico). $\chi^2_{\alpha,gl} = \chi^2_{0.05,\,(k-1)(f-1)}$

$\chi^2_{\alpha,gl} = \chi^2_{0.05,\,(2-1)(1-1)} = 1$, el valor en la tabla de χ^2_α, es 3.84

Establecer estadístico muestral (valor calculado) y calcular las frecuencias esperadas.

$$E_{i,j} = \frac{\sum Ok_{i,j} \left(\sum Of_{i,j}\right)}{N}$$

$$\chi^2_{calculado} = \sum_{j=1}^{f} \sum_{i=1}^{k} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$$

	Cazadores proveedores	**Cazadores club**	**Total**
Cazadores con éxito	$\frac{41*83}{118} = 28.83$	$\frac{41*35}{118} = 12.16$	41
cazadore con las manos vacias	$\frac{77*83}{118} = 54.16$	$\frac{41*35}{118} = 22.83$	77
Total	83	35	118

$$\chi^2_{calculado} = \frac{(26-28.83)^2}{28.83} + \frac{(57-54.16)^2}{54.16} + \frac{(15-12.16)^2}{12.16} + \frac{(20-22.83)^2}{22.83} = 1.44$$

Conclusión

El valor de $\chi^2_{calculado}$, es inferior al valor límite χ^2_α, es 3.84, la hipótesis H_0 es aceptada por lo que el porcentaje de cazadores pertenecientes al club que regresaron con las manos vacías no es diferente del porcentaje de cazadores proveedores de caza y pesca.

5 Test no Paramétricos de Comparación de Muestras

Literatura sugerida:

Zar. J. H., 1999. Bostatistical Analysis (4 edición) Prentice Hall. Estados Unidos. 663 p. (Pag. 146-155).

Cuando las condiciones de normalidad no son respetadas o que los datos son de naturaleza semicuantitativa (escala de variación ordinal) se utilizan los test no paramétricos. Estos test no se basan en los parámetros de las distribuciones y por consiguiente tampoco en las muestras de estos. Por tanto no están sujetos a las leyes de distribución de los elementos de la población. Se utilizan sobre todo en muestras pequeñas. Algunos de los test no paramétricos para muestras independientes son:

- Test de medianas, test de Wilcoxon-Mann-Whitney y el test de Kolmogorov-Smirnov

y para muestras pareadas se utiliza:

- Test de signos y test de Wilcoxon apareado.

5.1. Comparación de Dos Muestras Independientes: Test de Medianas

Desarrollado por Moo (1950) y Westenberg (1948) verifica la hipótesis de igualdad de medianas de 2 poblaciones. Supone que la forma de la distribución es similar a las 2 poblaciones comparadas.

Si la hipótesis es verdad:

H_0: $Me_1 = Me_2 = Me$

H_1: $Me_1 \neq Me_2 \neq Me$

La mejor estimación de la *Me* es la que se obtiene calculando la *Me* de la muestra si agrupamos las 2 muestras de origen:

Me(1+2) por tanto debe ser igual al 50% de los elementos de la muestra de la variable hacia arriba y hacia abajo. Se construye entonces una tabla de contingencia utilizando frecuencias observadas y esto se calcula utilizando la ley χ^2 a 1 *gl*.

$$\chi^2_{me} = \frac{4(f_1 f_4 - f_2 f_3)^2}{n n_1 n_2}$$

f_1, f_2, f_3 y f_4 = frecuencias observadas

Reglas de decisión:

Si el valor observado χ^2 es inferior al valor crítico, la hipótesis principal es aceptada, si no es rechazada y la hipótesis alternativa es aceptada.

La clasificación de los elementos está en función de su muestra de origen y de su valor en relación a la mediana como se muestra en la siguiente tabla:

Posición de elementos	Muestra 1	Muestra 2	Total
No. de elementos $X>Me_{(1+2)}$	f_1	f_2	$n/2$
No. de elementos $X<Me_{(1+2)}$	f_3	f_4	$n/2$
Total	$n1$	$n2$	$n=n_1+n_2$

Ejemplo 30. Se grabó el canto emitido por un pájaro en su territorio (sólo machos, pues son territoriales), si se reproduce el sonido, se observa una respuesta inmediata de los machos con el objeto de proteger su territorio. Un investigador quiere corroborar si la respuesta del ave es la misma con el canto de un ave originaria de norte América y que se supone es una subespecie de esta. Se tienen las siguientes muestras:

Difusión del canto			
Rapidez de respuesta/ muestra	**Ave x**	**Subespecie**	**Total**
Rápida	17	2	19
Media	3	6	9
Lenta	3	8	11
Sin respuesta	1	11	12
Negeativa (alejamiento)	0	5	5
Total	24	32	56

Pregunta: ¿La difusión del canto de la subespecie provoca reacciones rápidas en el ave x?

Condiciones: Datos semicuantitativos clasificados sobre escala ordinal, por lo tanto es válido trabajar con el test de medianas.

Planteamiento de las hipótesis

H_0: La difusión del canto de la subespecie no provoca ninguna reacción en al ave

H_1: La difusión del canto de la subespecie provoca reacciones diferentes en al ave

$H_0: M_{e1} = M_{e2}$

$H_1: M_{e1} \neq M_{e2}$

Establecimiento del test : $\chi^2_{me} = \dfrac{4(f_1f_4 - f_2f_3)^2}{nn_1n_2}$

Como la variable es el sonido, se tienen entonces 5 escalas diferentes. La mediana sería 2.5, que correspondería entre la respuesta media y lenta, se llena la tabla de contingencia, se suman todos los valores que están por encima de la *Me* para la muestra 1 para obtener f_1 y f_3. Para la subespecie; los valores para obtener f_2 y f_4 se obtienen sumando los valores que están por debajo de la mediana incluyendo el valor de la mediana; es decir:

17+3=20 y corresponde a f_1

3+1= 4 y corresponde a f_3

Ejemplo 30. Continuación

6+2= 8 y corresponde a f_2

8+11+5= y corresponde a f_4

	Ave	Subespecie	Total
$x>m_e$	f_1=20	f_2=8	28
$X<M_e$	f_3=4	f_4=24	28
Total	24	32	56

$$\chi^2_{Me} = \frac{4(20*24 - 8*4)^2}{56*24*32} = 18.66$$

Conclusión

El valor calculado de χ^2_{Me} es > al valor crítico $\chi^2_{0.01}$= 6.63, las medianas son diferentes y por tanto la respuesta del ave muestra no es la misma que la de la subespecie.

5.2. Muestras Independientes (Mann-Whitney)

Busca verificar si los elementos de dos grupos clasificados por orden creciente sobre una misma escala ordinal, ocupan posiciones o rangos equivalentes, que nos permita ver si hay similitud entre las distribuciones. Este test se basa en la variable *U de Mann-Whitney*, consiste en primer lugar en clasificar los elementos de dos muestras por orden creciente o decreciente, después en calcular *U* y *U'* que corresponden al número de veces que un elemento del 2° grupo antecede a un elemento del 1er grupo y viceversa.

Caso de pequeñas muestras: n_1 y $n_2 < 20$

Sólo se aplica cuando hay muestras que tienen entre 9 y 20 elementos.

$$U = n_1 n_2 + \frac{n_1(n_{1+1})}{2} - R_1$$

$$U' = n_1 n_2 + \frac{n_2(n_{2+1})}{2} - R_2$$

$$U' = n_1 n_2 - U$$

$$U = n_1 n_2 - U'$$

La siguiente tabla muestra la manera de utilizar el test de acuerdo al orden o "ranqueo" de los datos; y la tabla 7 anexa presenta los valores de U_α.

	H_0= Grupo 1 ≥ Grupo 2 H_1= Grupo 1 < Grupo 2	H_0= Grupo 1 ≤ Grupo 2 H_1= Grupo 1 > Grupo 2
"Ranqueo" del valor más bajo al más alto	*U*	*U'*
"Ranqueo" del valor más alto al más bajo	*U'*	*U*

Ejemplo 31. Probar si hay deferencia entre la talla de estudiantes hombres y mujeres de una manera significativa (α= 0.05).

Hombres talla (cm)	Mujeres talla (cm)	Orden hombres "ranqueo"	Orden mujeres "ranqueo"
193	175	1	7
188	173	2	8
185	168	3	10
183	165	4	11
180	163	5	12
178		6	
170		9	
n_1=7	n_2=5	R_1=30	R_2=48

Planteamiento de las hipótesis.

H_0: La talla de estudiantes hombres es igual a la de las mujeres

H_1: La talla de estudiantes hombres es diferente a la de las mujeres

Establecer el estadístico muestral

$$U = n_1 n_2 + \frac{n_1(n_{1+1})}{2} - R_1$$

Resolución del test

$$U = (7)(5) + \frac{(7)(8)}{2} 30 = 33$$

$$U' = (7)(5) - 33 = 2$$

Establecer el estadístico teórico

$$U_{0.05}, (5,7) = 30$$

Conclusión

Como 33 es mayor a 30, se rechaza la H_0 y por tanto se admite que la talla de los estudiantes hombres es mayor que la de las mujeres.

Ejemplo 32. Se quiere probar la eficiencia de dos métodos de enseñanza, comparándose los resultados del número de palabras por minuto que escriben a máquina los estudiantes de un colegio particular con los de la escuela pública. Los resultados son los siguientes:

Colegio particular No. palabras/ minuto	Escuela pública No. palabras/ minuto	Colegio particular "rango"	Escuela pública "rango"
44	32	9	3.5
48	40	12	7
36	44	6	9
32	44	3.5	9
51	34	13	5
45	30	11	2
54	26	14	1
56			
n_1=8	n_2=7	R_1=83.5	R_2 = 36.5

Como el "rango u orden" se realizó del valor más bajo al más alto, y como los valores de n_1 son mayores a los de n_2, se utilizará *U'* como test estadístico.

Ejemplo 32. Continución

Planteamiento de las hipótesis.

H_0: Los estudiantes del colegio particular escriben el mismo número de palabras por minuto en máquina de escribir que los de la escuela pública

H_1: Los estudiantes del colegio particular escriben mayor número de palabras por minuto en máquina de escribir que los de la escuela pública

Establecer el estadístico teórico

$$U_{0.05}, (7,8)= 43$$

Establecer el estadístico muestral

$$U' = n_1 n_2 + \frac{n_1(n_{1+1})}{2} - R_2$$

$$U = (7)(8) + \frac{(7)(8)}{2} - 36.5 = 47.5$$

Conclusión

Como 47.5 es mayor a 43, se rechaza la H_0 y por tanto se admite que el número de palabras por minuto que escriben en máquina los estudiantes de escuela particular es mayor que las que escriben los estudiantes de la escuela pública.

5.3. Comparación de Muestras Pareadas (test de signos)

El test no paramétrico de comparación de medias de dos muestras pareadas se basa en el análisis de muestras cuantitativas existentes entre cada par de datos. El test no paramétrico se distingue por el hecho de que no toma en cuenta más que el signo de las diferencias. Admitiendo que las muestras provienen de la misma población (hipótesis principal) las diferencias (d) tienen entonces la misma probabilidad de ser positivas que negativas, de donde sí H_0 es verdad:

$$P (d > 0) = P (d < 0) = 0.5$$

En esta hipótesis, la probabilidad de que *Yi* diferencias sean positivas entre n diferencias con valor mayor a cero, se da por la ley binomial del parámetros n y $p = 0.5$

2.3.1. Caso de Grandes Muestras

Si el número de parejas de datos presentan una diferencia mayor a cero, igual o superior a 30, la ley binomial tiende hacia la normal de parámetros $\mu = np = 0.5\, n$ y $\sigma = \sqrt{npq} = \sqrt{0.25n}$. En esta caso y admitiendo que la hipótesis principal de igualdad de poblaciones de origen es verdad la variable auxiliar:

$$Z_s = \frac{Y - 0.5n}{0.5\sqrt{n}}$$

Obedece a una ley normal centrada y reducida y representa el número de diferencias positivas y *n* el número de diferencias mayores a 0. Según la formulación de la hipótesis alternativa se efectuará entonces un test unilateral o bilateral.

Si H_1 declara que la primera población es más grande que la segunda, esto corresponderá:

$H_1 : P(d > 0) > 0.5$

H_0 será aceptada si Z_s.obs es inferior a Z_α

Si H_1 declara que la primera población es más pequeña que la segunda, esto correspondería:

$H_1 : P(d < 0) < 0.5$

H_0 será aceptada si Z_s.obs es superior a $-Z_\alpha$

Finalmente si H_1 declara que las dos poblaciones de origen son diferentes, esto correspondería:

$H_1 : P(d > 0) \neq 0.5$

H_0 será aceptado si Z_s.obs es inferior a $Z_{\alpha/2}$

5.3.2. Caso de Pequeñas Muestras

Si el número de pares de datos presentan una diferencia mayor a 0 e inferior a 30, la aproximación normal no es válida. Se tiene entonces que utilizar la distribución exacta de Y utilizando la tabla de probabilidades elementales de $P(Y_i)$ para P =0.5 y $n < 30$, Para deducir los valores críticos de Y, se tiene que partir de un extremo de la distribución y agregar los valores de $P(Y_i)$ hasta que se llegue a los valores de α o $\alpha_{/2}$ deseados a los que corresponden los valores críticos $Y_{\alpha_{/2}}$ buscados. En cada caso hay que satisfacer las expresiones siguientes:

$P(Y \geq Y_\alpha) = \alpha$ si la hipótesis alternativa corresponde a: H_1: $P(d > 0) > 0.5$

$P(Y \leq Y_\alpha) = \alpha$ si H_1 se escribe: $P(d > 0) < 0.5$

$P(Y \leq Y_{\alpha/2}) = \alpha/2$ y $P(Y \geq Y_{s\alpha/2}) = \alpha/2$ si H_1 se escribe: $P(d > 0) \neq 0.5$.

Las reglas de decisión dependen de la formulación de la hipótesis alternativa. Así la hipótesis principal será aceptada si:

$Y_{obs} < Y_o$ para H_1: $P(d > 0) < 0.5$

$Y_{obs} > Y_o$ para H_1: $P(d > 0) < 0.5$

$Y_{i_{\alpha/2}} > Y_{obs} <> Y_{s\alpha/2}$ para H_1: $P(d > 0) \neq 0.5$

Ejemplo 33. Refiriéndose al ejemplo del test de medianas, se quiere saber si un individuo dado de la subespecie de aves de Norteamérica, reacciona diferente a la difusión del canto de su propia subespecie y de la otra subespecie. Para cada individuo de la subespecie norte americana, se difunde primero el canto de su propia subespecie, después la de la otra subespecie y si se nota alguna reacción cuando se emiten estos sonidos, los resultados son semicuantitativos y se presentan en la siguiente tabla:

Rapidez de respuesta de la subespecie norteamericana cuando se emite el canto de la subespecie

Subespecie Norteamericana ***(T.t. hiemalis)***	**Subespecie europea** ***(T.t. troglodytes)***	**Número de aves**
Rápido	Rápido	2
Rápido	Medio	3
Rápido	Lento	3
Rápido	Ninguna	1
Rápido	Negativo	1
Medio	Medio	2
Medio	Lento	3
Medio	Ninguna	4
Medio	Negativa	2
Lento	Lento	1
Lento	Ninguna	3
Lento	Negativa	1
Ninguna	Ninguna	3
Ninguna	Negativa	1
Ninguna	Lento	1
Lento	Medio	1

Pregunta: ¿El ave de Norteamérica reacciona más cuando se emite el canto del ave europea?

Elección del test: las dos muestras son pareadas puesto que los cantos de las aves son difundidos a los mismos individuos; por otra parte, los datos son colectados sobre una escala de variación ordinal. Es por tanto necesario aplicar un test no paramétrico para muestras pareadas, el test de signos es uno de los métodos posibles.

Condiciones de aplicación: Ninguna condición de aplicación es requerida, sin embargo se tiene que aplicar un test para pequeñas muestras puesto que el número de diferencias son nulas e inferior a 30.

Planteamiento de hipótesis

H_0: La reacción al canto es parecida en ambas subespecies

H_1: la reacción al canto de la subespecie norteamericana es más rápida

H_0 : $P(d > 0) = 0.5$

H_1: $P(d > 0) > 0.5$ (test unilateral)

Test estadístico: el test estadístico para pequeñas muestras consiste en calcular el número Y_{obs} de aves que tengan una reacción más rápida ($d > 0$) cuando se reparte su propio canto.

Si H_0 es verdad, el número Y de elementos con $d > 0$ obedece a una ley binomial de parámetros P=0.5 y $n = 24$. El número n es igual a 24 puesto que de los 32 pares de datos, 8 tienen una diferencia negativa, que corresponde a reacciones idénticas de las dos difusiones del canto.

Reglas de decisión: Para encontrar el valor crítico Y_α, hay que satisfacer la expresión:

$$P(Y \geq Y\alpha) = \alpha$$

Ejemplo 33. Continuación

Al nivel de significatividad de 5%, la expresión se escribe:

$$P(Y \geq Y_{0.05}) = 0.05$$

Según la tabla correspondiente a la distribución binomial (Tabla 1), para $n = 24$ y $p = 0.05$;

Los valores de probabilidad se observan en la siguiente tabla:

Valores de probabilidad de la tabla de distribución binomial	Valores Acumulados
$P(24) = 0.000$	$P(Y = 24) = 0.000$
$P(23) = 0.000$	$P(Y \geq 23) = 0.000$
$P(21) = 0.000$	$P(Y \geq 21) = 0.000$
$P(20) = 0.001$	$P(Y \geq 20) = 0.001$
$P(19) = 0.003$	$P(Y \geq 19) = 0.004$
$P(18) = 0.008$	$P(Y \geq 18) = 0.012$
$P(17) = 0.021$	$P(Y \geq 17) = 0.033$
$P(16) = 0.044$	$P(Y \geq 26) = 0.077$

Como el valor acumulado de la probabilidad para $n = 24$ y $x = 16$ es 0.077 y este es mayor a 0.05, el valor crítico es por tanto 17, puesto que $P(Y \geq 17) = 0.033$.

Por tanto si Y_{obs} es superior o igual a 17, la hipótesis alternativa es aceptada con un riesgo de error igual a 0.033.

Aplicación del test: Se tienen que contar las aves norteamericanas que tengan una reacción más viva a la difusión de su propio canto que el canto del ave europea ($d>0$) : $Y_{obs} = 22$.

Decisión estadística: La hipótesis alternativa es aceptada puesto que $Y_{obs} > Y\ 0.05$.

Como $P(Y \geq 21) = 0.000$, (22-1 = 21) la hipótesis alternativa será aceptada al nivel α 0.05 ya que 0.033 es mayor que 0.000.

Interpretación: la reacción del ave al canto de la especie subamericana es más rápida que el canto de su propia especie.

5.4 Muestras Dependientes no Paramétricas (Wilcoxon), Muestras Pareadas

Este test involucra el cálculo de las diferencias de las muestras pareadas, entonces toma en cuanta el "rango" de las diferencias absolutas desde el valor más bajo al más alto y fija el signo de cada diferencia de valores del "rango". Entonces la suma de los valores del rango con el signo positivo llamada $T+$ y la suma de los valores del rango con el signo negativo, llamados $T-$ son obtenidas y se someten a la prueba de valores de la tabla de T. Se rechaza H_0 si $T+$ o $T-$ es menor o igual al valor crítico $T_{\alpha(2)}$, n, de la tabla de T (tabla 8 anexa).

Ejemplo 34. En un estudio morfométrico se quiso probar si existían diferencias significativas entre las longitudes de las extremidades delanteras con respecto a las traseras en venados, midiéndose 10 organismos y considerando un 5% de error. Se aplicó un prueba de Wilcoxon por rangos para definir si existían diferencias significativas entre ambas extremidades.

Los resultados se presentan en la tabla:

Venados	Longitud patas delanteras (cm) d_1	Longitud patas traseras (cm) d_2	Diferencia (d_1 - d_2)	Rango d_1	Rango d_2
1	142	138	4	4.5	4.5
2	140	136	4	4.5	4.5
3	144	147	-3	3	-3
4	144	139	5	7	7
5	142	143	-1	1	-1
6	146	141	5	7	7
7	149	143	6	9.5	9.5
8	150	145	5	7	7
9	142	136	6	9.5	9.5
10	148	146	2	2	2

Planteamiento de las hipótesis

H_0: La longitud de las extremidades delanteras y traseras de los venados no presentan diferencia

H_1: La longitud de las extremidades delanteras con respecto a las traseras de los venados es diferente

Establecer el estadístico teórico: $T_{0.05,\ (2),\ 10} = 8$

Establecer el estadístico muestral: $T+ = 4.5+4.5+7+7+9.5+7+9.5+2 = 51$

$T- = 3+1 = 4$

Conclusión

Como $T-$ es menor a $T_{0.05,\ (2),\ 10}$, se rechaza la H_0 y por tanto se admite que la longitud de las patas traseras de la muestra de venados es diferente a la longitud de las patas delanteras.

Entonces una vez $T+$ o $T-$, una u otra pueden obtenerse con las siguientes ecuaciones:

$$T- = \frac{n(n+1)}{2} - T+ \qquad \text{y} \qquad T+ = \frac{n(n+1)}{2} - T-$$

En el ejemplo anterior sería:

$$T- = \frac{10(10+1)}{2} - 51 = 4 \qquad T+ = \frac{10(10+1)}{2} - 4 = 51$$

Cuando se trabaja solamente con una cola de la distribución, los valores críticos $T+$ o $T-$ se deberán realizar como sigue de acuerdo a las hipótesis:

H_0: Las medidas de la población 1 ≤ a las medidas en la población 2

H_1: medidas de la población 1 > a las medidas en la población 2

Entonces H_0 es rechazada si $T- \leq T_{\alpha(1),n}$. para la hipótesis opuesta:

H_0: Las medidas de la población 1 ≥ a las medidas en la población 2

H_1: medidas de la población 1 < a las medidas en la población 2

Entonces H_0 es rechazada si $T+ \leq T_{\alpha(1),n.}$

6 Regresión y Correlación Simple

Literatura sugerida:

Scherrer B. 1984. Boestatistique. Gaëtan Morin editeur. 583-603 p.

Al analizar los datos para las ciencias, con frecuencia resulta que es conveniente saber la relación entre 2 variables. Por ejemplo, la relación entre la presión de la sangre y la edad, o la estatura y el peso, o entre la concentración de un medicamento inyectado y la rapidez de los latidos del corazón. El consumo de un nutriente y su ganancia en peso; la intensidad de un estímulo y el tiempo de reacción o hasta el ingreso total familiar y los gastos médicos. La naturaleza y la intensidad de las relaciones entre variables como estas pueden examinarse por medio del análisis de regresión y correlación.

La regresión es útil para averiguar la forma probable de la relación entre 2 variables y el objetivo es predecir o estimar el valor de una variable correspondiente a un valor dado de otra variable. El análisis de correlación se refiere a la medición de la intensidad de la relación entre las variables. Cuando se calculan medidas de correlación a partir de un conjunto de datos, el interés se centra en el grado de correlación entre las variables.

6.1. La Correlación

El concepto fue utilizado por primera vez en 1888 por el Sir Francis Galton.

Esta noción **mide el grado de unión** que hay entre varias variables, según la **naturaleza** y el **número** de variables implicadas se le asigna un nombre.

- La unión entre dos variables cuantitativas distribuidas normalmente, se denomina correlación **lineal simple**.
- La unión entre una variable dependiente y varias variables cuantitativas independientes, se denomina **correlación múltiple**.
- La unión entre dos conjuntos de variables cuantitativas, se denomina **correlación canónica**.
- La relación entre dos variables semicuantitativas, se denomina **correlación de rango**.
- La relación entre dos variables cualitativas, se denomina **asociación**; y se trata además de variables cualitativas binarias, se denomina coeficiente de **correlación de punto**.

6.1.1. La Correlación entre Dos Variables Cuantitativas (Pearson)

La correlación de Pearson es una medida de unión lineal existente entre dos variables cuantitativas aleatorias.

Esta dada por la siguiente expresión: $$\rho_{x,y} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

Donde: σ_{xy} es la covarianza entre x e y

$$\sigma_{xy} = \frac{\sum_{i=1}^{N} (x - \mu_x)(y - \mu_y)}{N}$$

σ_x y σ_y corresponden a las desviaciones estándar, estos términos se refieren a la población estadística.

Para una muestra aleatoria simple de talla "n," el estimador de $\rho_{x,y}$ es:

$$r_{xy} = \frac{S_{x,y}}{S_x S_y} = \frac{\sum(x - \bar{x})(y - \bar{y})}{(\sum(x - \bar{x})^2 \sum(\bar{y} - y)^2)^{1/2}}$$

Donde $S_{x,y}$ es la covarianza estimada:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x - \bar{x})(y - \bar{y})}{n - 1} = Sxy = \frac{n\sum(xy - (\sum x)(\sum y)}{n(n - 1)}$$

$$S_x^2 = \frac{\sum x2 - \frac{(\sum x)^2}{n}}{(n - 1)} = \text{varianza de x; } \sqrt{S_x^2} = \text{desviación estándar de x}$$

La correlación lineal es la covarianza de dos variables centradas y reducidas.

Las correlaciones pueden ser integradas en una tabla de doble entrada:

	x	y
x	$r_{x,x} = 1$	$r_{x,y}$
y	$r_{x,y}$	$r_{x,x} = 1$

Esta tabla presenta las siguientes propiedades (figura 37):

- Tiene solo 1 en la diagonal, ya que por definición la varianza de una variable centrada y reducida es 1; además de tener simetría respecto a la diagonal puesto que, $r_{x,y} = r_{x,y}$
- Varía entre -1 y +1, si $r = -1$ ó $+1$ todos los puntos están situados en una línea; si la nube de puntos no muestra ninguna tendencia.
- Y la última propiedad, es referente al signo. Si es (+) indica que la variable dependiente aumenta al mismo tiempo que la independiente. Si el signo es (-) significa que una variable aumenta cuando la otra disminuye.

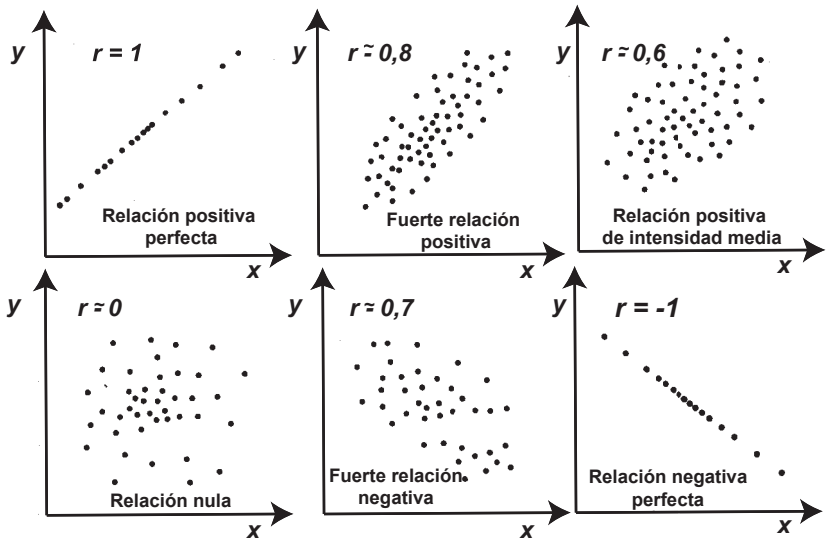


Figura 37. Coeficientes de correlación relacionados a diferentes nubes de dispersión; tomado de Scherrer (1984) (p. 585)

El coeficiente de correlación de Pearson, mide la intensidad de la unión y la eficacia de ajuste de los datos a un modelo lineal o linealizado; sin embargo, **no indica necesariamente una dependencia directa** de las variables o una relación causa-efecto.

Ejemplo 34. En un estudio de la capacidad reproductiva del insecto "A" del pino. Se busca determinar la intensidad de la relación entre la longitud del nido con el número de ovocitos.

El número de ovocitos por nido (y) y la longitud del nido (x) son dos variables aleatorias cuantitativas. Entonces su relación se mide por *r* de Pearson.

$S_{xy} = 5.51$

$$r_{xy} = \frac{S_{x,y}}{S_{x,}S_y} = \frac{\sum(x - \bar{x})(y - \bar{y})}{(\sum(x - \bar{x})^2 \sum(y - \bar{y})^2)^{1/2}}$$

$S^2_x = 0.3039$

$$r_{xy} = \frac{\sum_{i=1}^{n}(x - \bar{x})(y - \bar{y})}{n - 1} = Sxy = \frac{n\sum(xy - (\sum x)(\sum y)}{n(n - 1)}$$

$S^2_x = 344.13$

$$r = \frac{5.51}{\sqrt{0.3039} \bullet \sqrt{344.13}} = 0.544$$

La relación que une estas variables es débil. En otras palabras hay una fuerte dispersión de puntos.

6.1.2. Cálculo de Significatividad de *r*

Tomando como ejemplo el ejercicio anterior, se quiere probar si la correlación entre la longitud del nido y el número de ovocitos es altamente significativa. Para probar la significatividad de *r*, se utiliza la prueba de *t* con *n-2 grados de libertad*.

- Condiciones de aplicación del test:

  * Variables aleatorias cuantitativas,

  * con distribución normal.

  * Para que t_r sea válido, *x* e *y* deben tener una distribución binormal.

- H_0: $\rho=0$
  H_1: $\rho>0$ Prueba: $t_r = \frac{r\sqrt{n-2}}{\sqrt{1 - r^2}}$

r= 0.544 $t_r = \frac{0.544\sqrt{69-2}}{\sqrt{1-0.544^2}} = 5.24$

n= 69; α=0.01

Como n = 69 y no hay un valor preciso en la tabla y para test unilateral, el valor crítico se obtiene interpolando los valores de *gl= 65* y *gl*=70, entonces:

$t_{0.01(1),65}= 2.381= c_a$ C_a = Valor crítico de a

$t_{0.01(1),70}= 2.386= c_b$ C_b= Valor crítico de b

Se tiene entonces que *gl* a< *gl* b;

Ahora se estima la proporción utilizando n-2 gl; así se tiene: EQ

$$p= \frac{gl - a}{b - a} = \frac{67 - 65}{70 - 65} = \frac{2}{5} = 0.4$$

Se calcula el valor crítico: $c_{gl=67}=c_a+p(c_a-c_b)$=2.381+0.4*0.005=2.383

$$t_{0.01(1),67}= 2.383$$

Decisión estadística:

t_r=5.24 > t_α= 2.383 se rechaza H_0 , al 99% de confiabilidad, lo que indica que el número de ovocitos aumenta de manera muy significativa con la talla del nido.

6.1.3. La Regresión Lineal

La regresión lineal tiene un triple objetivo

- Permite **resumir** o **sintetizar** la relación existente entre una variable aleatoria dependiente "*y*" y una o varias variables aleatorias (**modelo II**) o controladas (**modelo I**) x_i llamadas variables explicativas (independientes).

- **Describe** la forma de relación que une a las variables. Puede ser una relación lineal; puede ser una asociación de 2°, 3° ó *n* grado, entonces se habla de una regresión polinomial, que pueden ser funciones hiperbólicas, logísticas u otras.

- **Predecir**; los valores de y_i en función de x_i. Es decir, estimar con un mínimo de error el valor desconocido de "*y*" de un elemento a partir de los resultados obtenidos de las variables predictivas x_i .

Cuando la estimación se realiza sobre varias variables descriptivas cuantitativas se trata de **regresión múltiple**. Si se agregan varias variables cualitativas, se trata de **regresión múltiple con variables mudas**.

Si sólo hay dos variables cuantitativas se trata de **regresión simple**.

6.1.4 Regresión Lineal Simple

Es una función de primer grado que une las variables *x* e *y*: $y=ax + b$

Se aplica a los modelos I y II, Corresponde a la recta que atraviesa de la mejor manera la nube de puntos cuando se relacionan dos variables.

6.1.5 Cálculo de la Recta de Regresión y Función de X: Método de Mínimos Cuadrados (MC)

Consiste en escoger una pendiente (*a*) y una ordenada al origen (*b*) de la recta que minimiza la suma de los cuadrados de los errores (SCE).

Error = Separación entre el valor observado y_i y el valor predicho por la recta y_i, es el ei residuo (figura 38).

El principio de los mínimos cuadrados es minimizar los errores:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

En cálculo diferencial se vió que el mínimo de una función se encuentra igualando a cero su primera derivada.

Como: $\hat{y}_i = ax_i + b$ *entonces:* $\sum_{i=1}^{n} e^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - (ax_i + b))^2$

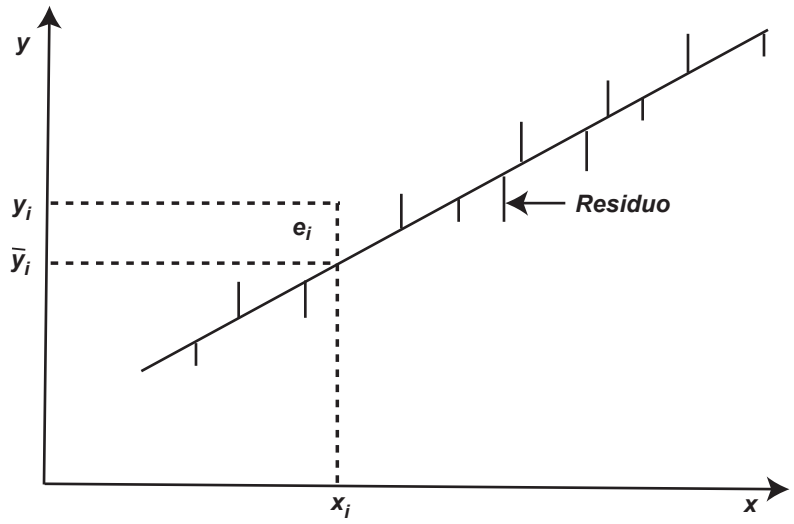


Figura 38. Gráfica de los residuos de una recta de regresión de y en x; tomado de Scherrer (1984) (pág, 626)

Se calculan las derivadas parciales en función de *a* y *b* y se seleccionan los parámetros o valores que satisfacen simultáneamente la anulación de las dos derivadas parciales.

Las dos ecuaciones con 2 incógnitas se llaman ecuaciones normales:

$$Sea: \quad S=\sum^{n} e_i^2; \qquad \frac{\partial s}{\partial b} = -2\sum(y_i - b - ax_i) = 0$$

$$\frac{\partial s}{\partial a} = -2\sum^{n} x_i\,(y_i - b - ax_i) = 0$$

Del desarrollo se obtiene:

$$b= \bar{y} - \overline{ax}\ ; \qquad a = \frac{S_{xy}}{S_x^2}$$

Ejemplo 35. Se quiere obtener el modelo de regresión lineal de la concentración de DDt, DDE, DDD contra la edad de algunos peces:

n=45	$\bar{x}$=3.44 años
$\sum x$=155	S_x= 1.235
$\sum x^2$= 601	S^2_x=1.525

Modelo I	
$\sum y$=20.09	*Sy= 0.276*
$\sum y^2$= 12.33	$\bar{y}$=0.446µg/g•año
$\sum xy$=83.325	S^2_x=0.076

$$S_{xy}=\frac{n\sum xy - (\sum x)(\sum y)}{n(n-1)}\ ;$$

$$S_{xy}=\frac{45x83.32-155x20.09}{45(44)} = 0.321$$

$$r=\frac{S_{xy}}{S_x S_y}=\frac{0.321}{1.235 \times 0.276} = 0.94\ ; \quad a=\frac{S_{xy}}{S^2_x}=\frac{0.321}{1.525}= 0.210\text{µg/g • año}$$

Ejemplo 35. Continuación

$b = \bar{y} - a\bar{x} = 0.446 - 0.210 \times 3.44 = -0.278 \mu g$; $y = 0.210x + (-0.278)$ o bien

$y = -0.278 + 0.210x$

Ejemplo 36. Obtener el modelo de regresión lineal entre el número de huevos por la longitud del nido (Modelo II, porque las dos son aleatorias):

Longitud del nido	No. ovocitos
$\bar{x}$= 8.74 mm	$\bar{y}$= 59.96 ovocitos
S^2_x = 0.3039 mm^2	S^2_y=344.13 ovocitos^2
S_{xy}= 5.51	r= 0.539

$$a = \frac{S_{xy}}{S^2_x} = \frac{5.51}{0.3039} = 18.13 \text{ ovocitos/mm}$$

$b = \bar{y} - a\bar{x} = 59.96 - 18.13 \times 8.74 = -98.5$ ovocitos

$y = 18.13x - 98.50$

6.2 Coeficiente de Determinación r^2

Mide la proporción de la variación de *y* explicada por la variación de *x* Para la regresión lineal simple:

$$R^2 = r^2$$

La variación puede ser dividida como sigue:

Dispersión total de y = dispersión debido a los errores + dispersión debida a la regresión

$$\sum(y_i - \bar{y})^2 = \sum(y_i - \hat{y}_i)^2 + \sum(\hat{y}_i - y)^2$$

$$CET = SCE + SCER$$

Entonces: La dispersión de *y* en relación a la media, es igual a la dispersión respecto al valor predicho (calculado) más la dispersión del valor predicho respecto a la media. La figura 39 y 40 muestra la descomposición de la dispersión total de *y* en sus dos componentes. 1) dispersión debido a los errores, 2) dispersión debido a la regresión, 3) disperción total. 3=2+1.

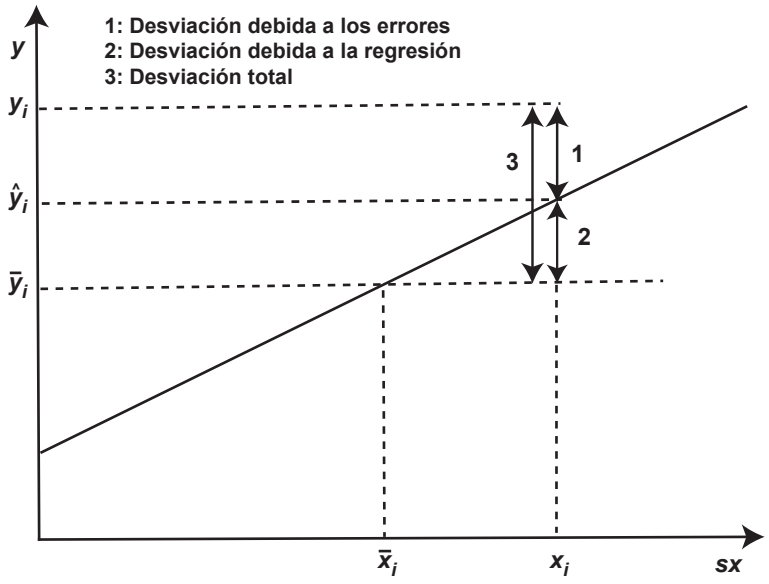


Figura 39. Dispersión total y sus componentes; tomado de Scherrer (1984) (p. 636)

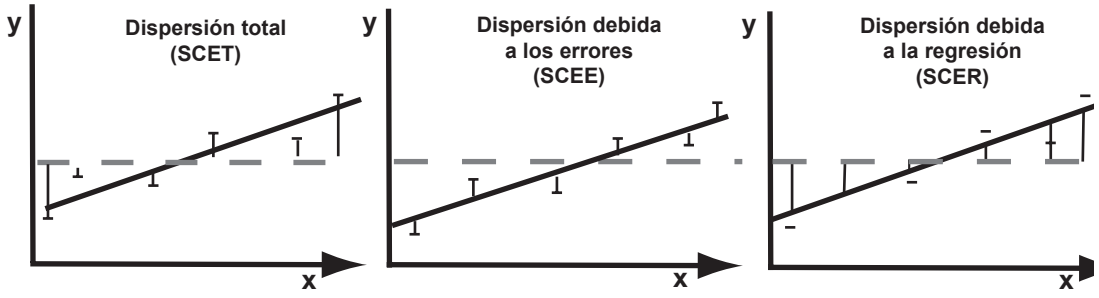


Figura 40. Descomposición de la dispersión total en sus dos componentes: SCET=SCE+SCER; tomado de Scherrer (1984) (p. 637)

$$r2 = \frac{SCER}{SCET} = \text{Dispersión explicada por regresión/dispersión total}$$

$$r^2 = \frac{S^2_{xy}}{S^2_x S^2_y} = r^2$$

Si todos los puntos están alineados, e_i son nulos (SCE=0);

$$SCET=SCER \; ; \; r^2=1$$

Ejemplo 37. Capacidad de reproducción. retomando el ejemplo anterior se tiene:

$$r = 0.539 \quad ; r^2 = 0.29$$

Así, 29% de la variación de ovocitos se explica por la variación del nido como la relación biunívoca, pues se trata de un modelo II, 29% de la variación de la longitud del nido es explicada por la variación del número de ovocitos.

Los valores utilizados son:

x= 8.74 mm; $\bar{y}$= 59.96 ovocitos

s^2_x=0.3039 mm^2; S^2_y= 344.13 ovocitos^2

S_{xy}= 5.51 ; r= 0.539

$$a = \frac{S_{xy}}{S^2_x} = \frac{5.51}{0.3039} = 18.13 \text{ ovocitos/mm;}$$

$b = \bar{y} + a\bar{x}$ = 59.96 - 18.13 * 8.74= 98.50 ovocitos;

La ecuación se escribe: y=18.13*x-98.50

6.3 Significatividad de la Regresión de *Y* en *X*

Esta consiste en verificar o probar si R^2 o **a** son significativamente diferentes de cero.

Se basa en el principio de análisis de varianza o el de la distribución de muestreo de la pendiente. Para ambos casos se tiene que:

H_0: α=0; o H_0: ρ^2=0;

H_1: α y ρ^2 son diferentes de cero

Para verificar si a=0, se supone que: Y es una variable aleatoria y X un factor controlado (modelo I). Y está unido a X_i por la expresión:

$$Y_i = \alpha X_i + \beta + \varepsilon_i; \quad con\ i = 1,2,3,\ldots,n$$

El error ε_i, para la población es denominado e_i para la muestra, es una variable aleatoria cuya esperanza matemática es cero y cuya varianza es σ_2.

$$E(\varepsilon_i) = 0; \qquad Var\ (\varepsilon_i) = \sigma_2$$

Y_i tiene dos componentes: uno cierto dado por $\alpha X + \beta_i$; otro aleatorio ε_i.

En resumen las varianzas de Y_i y de ε_i, son constantes para cualquier valor de X.

6.3.1 Prueba de Significatividad de la Pendiente *a*

Prueba por t de $t_{ac} = \frac{a}{\sqrt{var(a)}}$ *Student:* ; t_{ac} *calculado*

Entonces: $var(a) = \frac{S_e^2}{(n-1)\,S_x^2} = \frac{SCE}{(n-2)(N-1)\,S_x^2}$

$$t_{ac} = \frac{a}{\sqrt{\frac{S_e^2}{(n-1)\,S_x^2}}}$$

Ejemplo 38. Retomando los valores del ejemplo anterior de reproducción, además de la tabla original.

Prueba de hipótesis:

Pregunta: ¿La regresión de *y* en *x* es significativa?

Condiciones: Dos condiciones: la normalidad y la equivarianza de las distribuciones de la variable y o de los errores e_i que corresponden a los diferentes valores de *x*. Un estudio los residuos permitirá comprobarlo.

Hipótesis: H_0: a = 0; H_1: a ≠ 0 prueba bilateral.

Prueba estadística: $t_{ac} = \frac{a}{\sqrt{\frac{S_e^2}{(n-1)\,S_x^2}}}$

Reglas de decisión. Para α= 0.05 y *gl*= *n*-2 (69-2=67) grados de libertad. Es de notar que *n*-2 se debe a que hay 2 parámetros a estimar (*a* y *b*). El valor crítico de $t_{\alpha/2,0.05}$. En consecuencia, H_0 es rechazada si $|t_{ac}| > t_{\alpha/2,0.05}$.

Cálculo: $S_e^2 = \frac{\sum (y_i - \hat{y})^2}{n-2} = \frac{SCE}{n-2} = \frac{16602}{67} = 247.8$

$S_x^2 = 0.3039$

$$t_{ac} = \frac{18.13}{\sqrt{\frac{247.83}{0.3039*68}}} = \frac{18.13}{67} = 5.24$$

Decisión estadística. t_{ac}=5.24 > $t_{\alpha/2,0.05}$. Entonces H_0 se rechaza al nivel de α=0.05.

Conclusión biológica. El número de ovocitos por nido está unido positivamente a la longitud del nido.

6.3.2 Análisis de Varianza: Prueba de Significatividad de r^2

Este enfoque consiste en separar la varianza total de *y* en sus dos componentes: la varianza explicada por la regresión y en consecuencia por la variable *x*, y la varianza debida a una multitud de factores no estudiadas en el modelo de regresión, que se denomina varianza residual o debida a los errores.

Análisis de varianza para verificar la significatividad de *a* y r^2

Fuente de variación	Suma de cuadrados de las desviaciones SCE	Grados de libertad	Varianza
Total	$SCET= \sum(y_i - y_i)^2$ $SCET = \sum y^2_i - \frac{(\sum y_i)^2}{n}$	n-1	S^2_y
Explicada por la regresión	$SCER= \sum (\hat{y}_i - \bar{y})^2$ $SCER=a^2 \sum(x_i - \bar{x})^2$ $SCER= r^2\sum(y_i - \bar{y})^2$ $SCER= r^2 * SCET$	1	$S^2_{yR}=SCER$
Residual o debida a los errores	$SCER= \sum(y_i - \hat{y})^2$ $SCE= SCET - SCER$	n -2	$S^2_e= \frac{SCE}{n-2}$

Hipótesis:

H_0: ρ^2=0;

H_1: ρ^2 son diferentes de cero

Prueba estadística:

$$Fc = \frac{S^2_{yR}}{S^2_e} = \frac{SCER}{\frac{SCE}{(n-2)}} = \frac{r^2 SCER}{\frac{SCE}{n-2}} = Fc = \frac{r^2 (n-2)}{1 - r^2} = Fc = \frac{a^2}{var(a)}$$

Es de notar que esta expresión es igual al cuadrado de $t_r = \sqrt{\frac{r^2(n-2)}{1- r^2}}$;

de esta forma este análisis se puede llevar a cabo por un análisis de varianza (prueba de *F*), o por una prueba de t_r. Generalmente se prefiere el de *F* que es “unilateral”.

Ejemplo 39. Sobre la capacidad reproductiva número de ovocitos en función de la talla del nido.

Hipótesis:

H_0: $\rho^2=0$;

H_1: $\rho^2 \neq 0$

Prueba:

$$Fc = \frac{S^2_{yR}}{S^2_e} = \frac{SCER}{\frac{SCE}{(n-2)}}$$

Reglas de decisión. Contrariamente al test de significatividad de la pendiente, el análisis de varianza es siempre unilateral: cuando la pendiente es positiva, se verifica si *a* o *r* son más grandes que cero; si es negativa, se verifica si son más pequeños que cero. Para $gl_1=1$, y $gl_2= n - 2= 67$ y $\alpha= 0.05$, se tiene: $F_{0.05,\ 1,67}= 3.96$ a $4.00 \approx 4.00$. Si $Fc \geq F_{0.05,1,67}$, se rechaza H_0.

Cálculo:

$$Fc=\frac{S^2_{yR}}{S^2_e} = \frac{SCER}{\frac{SCE}{(n-2)}} = \frac{6792.5}{247.8} =270.4$$

Análisis de varianza para verificar la significatividad

Fuente de variación	Suma de cuadrados de las desviaciones SCE	Grados de libertad	Varianza
Total	$SCET= \sum(y_i - y_i)^2$ $SCET = \sum y^2_i - \frac{(\sum y)^2}{n}$	n-1=68	S^2_y=344.13
Explicada por la regresión	$SCER= \sum (\hat{y}i - \bar{y})^2$ $SCER=a^2 \sum(x_i - \bar{x})^2$ $SCER= r^2\sum(y_i - \bar{y})^2$ $SCER= r^2 * SCET$	1	$S^2_{yR}=SCER$=6792.5
Residual o debida a los errores	$SCER= \sum(y_i - \hat{y})^2$ $SCE= SCET - SCER$= 16608	n -2=67	$S^2_e= \frac{SCE}{n-2}$= 247.8

Decisión estadística: Se rechaza H_0.

Conclusión: El número de ovocitos por nido, está significativamente relacionado con la talla del nido.

6.3.3 Intervalo de Confianza de la Pendiente de una Recta de Regresión de *y* en *x*

Está dado por: $Pr\left[a - t_{\alpha/2,n-2}\sqrt{var(a)} < \alpha < a+t_{\alpha/2,n-2}\sqrt{var(a)}\right] = 1 - \alpha$

Donde : $var(a)=\dfrac{S^2_e}{(n-1)S^2_x}$

Ejemplo 40. Del ejemplo de capacidad reproductora, buscar el intervalo de confianza de la pendiente.

$var(a)=\dfrac{S^2_e}{(n-1)\,S^2_x}$; $\qquad S^2_e=\dfrac{\sum(y_i-\hat{y}_i)^2}{n-2}=\dfrac{SCE}{n-2}=\dfrac{16602}{67}=247.79$

$S^2_x = 0.3039$

$var(a)=\dfrac{247.79}{68 * 0.3039} = 11.97$

$t_{\alpha/2,67}= 2.0$; $S_2=\sqrt{11.97} = 3.46$; $a= 18.13$

$Pr\,[18.13 - 2 * 3.46 < \alpha < 18.13 + 2 * 3.46]= \alpha - 1$

$Pr\,[11.21< \alpha <25.05]= 0.95$

6.3.4 Intervalo de Confianza de la Ordenada al Origen

Está dado por: $Pr\left[b - t_{\alpha/2,n-2}\sqrt{var(b)} < \beta < b+t_{\alpha/2,n-2}\sqrt{var(b)}\right]= 1 - \alpha$

Donde: $var(b) =\dfrac{S^2_e\sum x^2_i}{n\sum(x_i-\bar{x})^2}$

Ejemplo 41. Retomando el ejemplo de la capacidad reproductora.

$\sum x^2_i = 5302.75$

$\sum (x_i - \bar{x})^2 = 74.0375$ $\quad var(b) = \frac{S^2_e \sum x^2_i}{n\sum (x_i - \bar{x})^2} = \frac{247.8 * 5302.75}{69 * 20.813} = 914982$

$\bar{x} = 8.75$ S^2_e

$S_b = \sqrt{914.982} = 30.249$

$S^2_e = \frac{\sum (y_i - \hat{y}_i)^2}{n - 2} = \frac{SCEE}{n - 2} = \frac{16602}{67} = 247.79$

$Pr[-98.5 - (2*30.249) < \beta < -98.5 + (2*30.249)] = 1 - \alpha$

$Pr[-158.998 < \beta < -38.002] = 0.95$

Valores observados de talla nido (mm) y de número de ovocitos

No. de nodos	Talla nido mm	No. ovocitos	No. de nidos	Talla nido mm	No. ovocitos
n	*X*	*Y*	*n*	*X*	*Y*
1	8.5	60	36	9	78
2	8	27	37	8.5	66
3	9.2	72	38	9	71
4	7.7	41	39	9.2	67
5	8.5	66	40	8.8	85
6	8	46	41	7.8	48
7	9.1	57	42	8.7	49
8	9	99	43	9	39
9	9.3	85	44	9.3	76
10	8.4	48	45	8.5	82
11	9.5	86	46	9.8	48
12	8.2	47	47	9.4	73
13	9.5	93	48	8.9	68
14	8.9	45	49	7.9	29
15	8.5	55	50	8.2	28
16	9.1	79	51	8.8	47
17	8.5	61	52	8	46
18	8.5	77	53	9	55
19	8.5	77	54	8.5	47
20	8.9	43	55	8.9	85
21	8.5	56	56	8.7	72
22	7.4	25	57	8.8	67
23	10	556	58	8.8	60
24	9.5	89	59	8.6	53
25	7.8	37	60	8.4	60
26	8.8	51	61	9.4	32
27	9.5	889	62	8.8	69
28	8.8	42	63	9.5	98
29	9	33	64	9	58
30	9.4	65	65	8	43
31	7.8	42	66	8.5	64
32	8.6	57	67	8.6	70
33	7.8	48	68	9.1	33
34	9.1	85	69	8.8	57
35	9.7	77			

6.3.5 Prueba de Hipótesis (significatividad) de *b*

Ejemplo 42. Sobre la capacidad reproductiva.

Prueba a efectuar: $t_{bc} = \frac{b}{\sqrt{var(b)}}$; donde: $var(b) = \frac{S^2_e \sum x^2_i}{n\sum (x_i - \bar{x})^2}$

Hipótesis. H_0: b=0; H_1= b ≠ 0 prueba bilateral

Reglas de decisión. Para un nivel de α=0.05, $t_{\alpha/2,v=n-2}$=2.0; entonces H_0 será rechazada si:

$$|t_{bc}| \geq t_{\alpha/2,v=n-2}$$

Cálculo

$\sum x^2_i$= 5302.75

$\sum (x_i - \bar{x})^2$ = 74.0375 $\quad var(b) = \frac{S^2_e \sum x^2_i}{n\sum (x_i - \bar{x})^2} = \frac{247.8 * 5302.75}{69 * 20.813}$ = 914.982

x= 8.75 $\quad S^2_e$= 247.79 $\quad S_b = \sqrt{914.982}$= 30.249

$t_{bc} = \frac{b}{\sqrt{var(b)}} = \frac{-98.5}{30.249}$ = -3.256; entonces, |-3.256| > $t_{0.05/2,67}$= 2.0,

por lo que se rechaza H_0. Esto se comprueba además por los intervalos de confianza calculados anteriormente, en los cuales el valor de b=0 está excluido de dichos intervalos.

6.3.6 Intervalo de Confianza para Estimaciones de $\hat{Y}_i$ y $\bar{\hat{Y}}_i$

Como a partir de X_i, se puede deducir con la ayuda del modelo de regresión, el valor estimado $\hat{Y}_i$ **que** le corresponde ($\bar{\hat{Y}}_i$), el objetivo de esta sección es de calcular a través del modelo de regresión el intervalo de confianza del valor estimado $\hat{Y}_i$ que corresponde a un valor x_i ; sucede también que se tenga que calcular el intervalo de confianza para la media de estimaciones $\bar{\hat{Y}}_i$, para el valor x_i . De esta manera el intervalo de confianza de la predicción de $\hat{Y}_i$ está dado por:

$$Pr[\hat{y}_i - t_{\alpha/2,n-2}\sqrt{var(\hat{y}_i)} < \mu_{yi} < \hat{y}_i + t_{\alpha/2,n-2}\sqrt{var(\hat{y}_i)}] = 1 - \alpha$$

Donde: $$var(\hat{y}_i) = S^2_e \left[\frac{1}{n} + \frac{(x_i + \bar{x})^2}{\sum (x_i - \bar{x})^2}\right]$$

Así la figura 41, muestra los intervalos de confianza de $\hat{Y}$ para el ejemplo sobre la talla del nido y el número de ovocitos. Se puede constatar que esta varianza es mínima cuando $x_i = \bar{x}$. Por otra parte, se hace cada vez más grande en la medida en que x_i se aleja de la media $\bar{x}$; por lo tanto, los límites de confianza se alejan de la recta de regresión cuando x_i se aleja de $\bar{x}$.

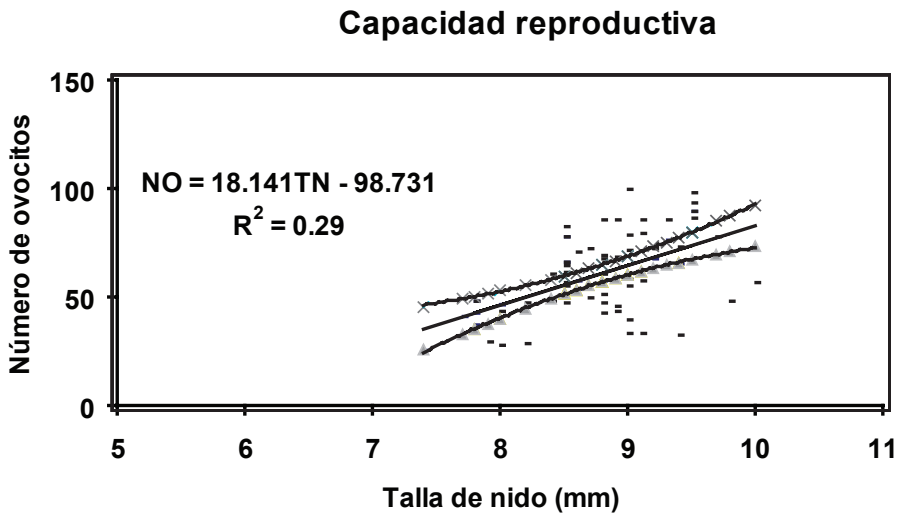


Figura 41. Representación de los límites de confianza de $\hat{Y}_i$.

Finalmente el Intervalo de confianza de la predicción $\hat{Y}_i$ está dado por:

$$Pr[\hat{y}_i - t_{\alpha/2,n-2}\sqrt{var(\hat{Y}_i)} < \mu_{yi} < \hat{y}_i + t_{\alpha/2,n-2}\sqrt{var(\hat{Y}_i)}] = 1$$

En este caso, la media $\hat{Y}_i$ es estimada a partir de *m* elementos que tienen el mismo valor x_i.

Donde: $$var(\hat{y}_i) = S^2_e\left[\frac{1}{n} + \frac{(x_i + \bar{x})^2}{\sum(x_i - \bar{x})^2}\right]$$

6.4. Coeficiente de Correlación de Spearman

Literatura sugerida:

http://es.wikipedia.org/wiki/Coeficiente_de_correlaci%C3%B3n_de_Spearman

http://www.fortunecity.com/campus/lawns/380/estadistica/coeficientecsrs.htm

Esta prueba estadística permite medir la correlación o asociación de dos variables y es aplicable cuando las mediciones se realizan en una escala ordinal, aprovechando la clasificación por rangos. El Coeficiente de correlación de Spearman ρ (rho), es una prueba no paramétrica que mide la asociación entre dos variables discretas. Para calcular ρ, los datos son ordenados y reemplazados por su respectivo orden.

El estadístico ρ viene dado por la expresión:

$$\rho = 1 - \frac{6\sum D^2}{N(N^2 - 1)}$$

donde *D* es la diferencia entre los correspondientes valores de *x - y*. *N* es el número de parejas.

Se tiene que considerar la existencia de datos idénticos a la hora de ordenarlos, aunque si estos son pocos, se puede ignorar tal circunstancia.

Para muestras mayores de 20 observaciones, podemos utilizar la siguiente aproximación a la *distribución de t de Student.*

$$t = \frac{\rho}{\sqrt{(1 - \rho^2) / (n - 2)}}$$

El coeficiente de correlación de Spearman se rige por las reglas de la correlación simple de Pearson, y las mediciones de este índice corresponden de + 1 a - 1, pasando por el cero, donde este último significa no correlación entre las variables estudiadas, mientras que los dos primeros denotan la correlación máxima.

La ecuación utilizada en este procedimiento, cuando en el ordenamiento de los rangos de las observaciones no hay datos empatados o ligados, es la siguiente:

$$r_s = 1 - \frac{6 \sum d^2}{N^3 - N}$$

Donde:

r_s = coeficiente de correlación de Spearman.

d^2 = diferencias existentes entre los rangos de las dos variables, elevadas al cuadrado.

N = tamaño de la muestra expresada en parejas de rangos de las variables.

Pasos.

1. Clasificar en rangos cada medición de las observaciones.
2. Obtener las diferencias de las parejas de rangos de las variables estudiadas y elevadas al cuadrado.
3. Efectuar la sumatoria de todas las diferencias al cuadrado.
4. Aplicar la ecuación.
5. Calcular los grados de libertad (gl). gl = número de parejas - 1. Solo se utilizará cuando la muestra sea mayor a 10.
6. Comparar el valor *r* calculado con respecto a los valores críticos de la tabla de valores críticos de coeficiente de correlación por rangos de Sperman en función de probabilidad (Tabla 8 anexa).
7. Decidir si se acepta o rechaza la hipótesis.

Ejemplo 43. Un investigador está interesado en conocer si el desarrollo mental de un niño esta asociado a la educación formal de su madre. De esta manera, obtiene la calificación de desarrollo mental en la escala de Gesell de ocho niños elegidos aleatoriamente y se informa del grado de escolaridad de las madres.

Elección de la prueba estadística

Se desea medir asociación o correlación. Las calificaciones de la educación formal de las madres están dadas en una medición cualitativa, pero tienen una escala ordinal, por lo cual es posible ordenarlas en rangos.

Planteamiento de la hipótesis

• Hipótesis alternativa (H_1). El desarrollo mental de los hijos es una variable dependiente de la educación formal de la madre; por lo tanto, existe una correlación significativa.

• Hipótesis nula (H_0). La asociación entre las variables de educación formal de la madre y el desarrollo mental de los hijos no es significativa, ni hay correlación.

Nivel de significación. Para todo valor de probabilidad igual o menor que 0.05, se acepta H_1 y se rechaza H_0.

Zona de rechazo. Para todo valor de probabilidad mayor que 0.05, se acepta H_o y se rechaza H_1.

Aplicación de la prueba estadística. Las observaciones de cada variable se deben ordenar en rangos, así como obtener las diferencias entre los rangos, efectuar la sumatoria y elevar ésta al cuadrado.

Calculo de r_s de Spearman.

$$r_s = \frac{1 - 6\sum d^2}{N^3 - N} = \frac{1 - 6 \times 26}{8^3 - 8} = \frac{1 - 156}{504} = 0.69$$

Cálculo de los grados de libertad (*gl*).

gl = numero de parejas - 1 = 8 - 1 = 7

Desarrollo mental de algunos niños y escolaridad de las madres.

Escolaridad de la madre (*X*)	Calificación del desarrollo mental de los niños (*Y*)
Primero de secundaria	90
Primero de primaria	87
Profesionista	89
Sexto de primaria	80
tercero de primaria	85
Analfabeta	84
Preparatoria	91

Educación de algunas madres y calificación de desarrollo mental de los hijos.

Rango de la eduación materna	Rango del desarrollo mental del niño	d	d^2
5	7	-2	4
2	5	-3	9
8	6	2	4
4	2	2	4
6	4	2	4
3	3	0	0
1	1	0	0
7	8	-1	1
N= 8			$\sum d^2=26$

El valor r_s calculado se compara con los valores críticos de r_s del coeficiente de correlación por rangos de Spearman.

El valor crítico de r_s con 7 grados de libertad, para una probabilidad de 0.05 del nivel de significatividad es 0.714, o sea, mayor que el calculado. Por lo tanto, éste tiene una probabilidad mayor que 0.05.

Decisión. Como el valor de probabilidad de r_s de 0.69 es mayor que 0.05, se acepta H_0 y se rechaza H_1.

Ejemplo 43. Continuación

Interpretación. El coeficiente de correlación de Spearman de 0.69 es menor que los valores críticos de la tabla, pues a éstos corresponde la probabilidad de obtener esa magnitud, al nivel de confianza de 0.05 y 0.01, para 0.714 y 0.893. Esto significa que para aceptar H_1, se requiere tener un valor igual o más lato que 0.714. Por lo tanto se acepta H_0 y se rechaza H_1, aun cuando, como se observa en la figura 42, existe una asociación relativa entre la educación formal de la madre y el desarrollo mental de sus hijos; sin embargo, ésta no es significativa.

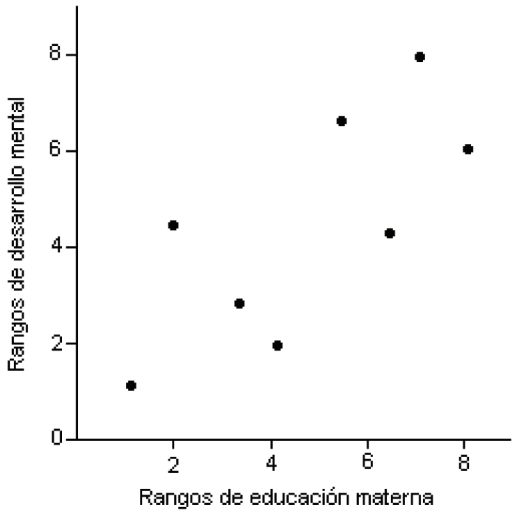


Figura 42. Asociación relativa entre la educación formal de la madre y el desarrollo mental de sus hijos tomado de: http://www.fortunecity.com/campus/lawns/380/estadistica/coeficientecsrs.htm

Para tener una mayor aproximación matemática de los interesados por los temas tratados en esta obra, se sugiere consultar cualquiera de los libros y páginas web indicadas en la literatura sugerida a continuación:

Lieteratura consultada:

Bonnier G y Tedin O. 1996. Bioestadística. Ed. Acribia. 223 p

Daniel W. W. 1982. Bioestadística. Limusa. 485 p.

King, B. M., Minium E. M., 2003, Statistical reasoning, 52-53.

Reyes Castañeda P. 1990. Estadística aplicada. Ed. Trillas. 216 p

Scherrer B. 1984. Boestatistique. Gaetan Morin editeur. 850 p

Sokal R. y F. J. Rholf. 2000. Biometry. Tercera edición. Ed. Freeman. 887 p

Zar J. H. 1999, Bioestatistical análisis. 4 edición. Pretince Hay. 663 p.

Sitios web:

http://euler.ciens.ucv.ve/pregrado/estadistica/archivos/guias-teo/guia1.pdf#search='estad%C3%ADstica%20descriptiva'

http://facultad.sagrado.edu/ConceptosBasicos.pdf#search=%22tabla%20de%20probabilidad%20conjunta%22

http://es.wikipedia.org/wiki/Funci%B3n_de_densidad

http://personal5.iddeo.es/ztt/Tem/t21_distribucion_normal.htm

http://www.bioestadistica.uma.es/libro/node38.htm

http://www.itcomitan.edu.mx/tutoriales/estadistica/contenido/unidad_4.html

ESTADÍSTICA I

Se realizó en el Departamento de Difusión y Publicaciones
del Centro EPOMEX-Universidad Autónoma de Campeche
Composición, diseño y proceso editorial a cargo de Jorge Gutiérrez Lara

Diciembre 2007

FACULTAD DE CIENCIAS
QUÍMICO BIOLÓGICAS

CENTRO
EPOMEX
UNIVERSIDAD AUTÓNOMA DE CAMPECHE