

BIO 149-P BIOMETRIA II

Clase 1. Introducción

INTRODUCCIÓN

En este curso vamos a ver una parte de la estadística aplicada que tiene que ver con la **estimación** (de parámetros) estadísticos y principalmente con la **prueba de hipótesis** estadísticas. La idea de este curso es construir sobre las ideas básicas de probabilidades, distribución de probabilidades, intervalos de confianza y prueba de hipótesis que ustedes vieron en el curso introductorio BIO-242A. Este curso, BIO-149P, parte del supuesto que ustedes tienen una noción general de qué son y cómo se calculan probabilidades de eventos bajo distintas condiciones y que además entienden bien el concepto de *incertidumbre* en el proceso de investigación científica. Puesto que algunos de ustedes necesitarán recordar algunos de los conceptos generales de prueba de hipótesis, en las primeras dos clases repasaremos brevemente los conceptos más importantes. Es recomendable que consulten sus apuntes del curso BIO-242A o un libro introductorio si este breve repaso no es suficiente para que se sientan seguros de que entienden bien estos conceptos.

En el curso veremos muy por encima lo que es el cálculo de incertidumbre en la estimación de parámetros a través de intervalos de confianza, para concentrarnos en el diseño de experimentos y la prueba de hipótesis. La poca atención que prestaremos en este curso a la estimación de parámetros no es un reflejo de que esta parte de la estadística sea menos importante que la prueba de hipótesis, sino que solamente un compromiso de tiempo. No solamente estaremos retringidos en el ámbito de la estadística que abordaremos, sino que además nos concentraremos en una técnica de análisis en particular, denominada **Análisis de Varianza**. Así, dejaremos de lado muchas otras aproximaciones estadísticas a la prueba de hipótesis (e.j. técnicas de distribución libre, pruebas de randomización, aproximación Bayesiana a prueba de hipótesis, etc.). En mi experiencia, es más provechoso aprender bien una o unas pocas técnicas estadísticas que el tener un curso lleno de “recetas” que son memorizadas y luego aplicadas en forma dudosa y muchas veces incorrecta.

Nuestra aproximación general corresponde a una gran rama de la estadística denominada “frecuentista”, la cual visualiza probabilidades de eventos como las frecuencias esperadas de ese evento particular, si el experimento o ejercicio pudiera repetirse de igual forma muchas (cientos, miles) veces. Aquí es solamente importante que entiendan que esta filosofía frecuentista no es la única manera de visualizar las probabilidades de eventos. Una rama importante de la estadística, aún más antigua que la estadística frecuentista, es la denominada estadística Bayesiana que se basa en el teorema de Bayes. La filosofía de la estadística Bayesiana visualiza probabilidades como el grado de confianza que nosotros podemos tener en que ocurra un determinado evento. La diferencia puede parecer sutil para ustedes, pero es suficientemente distinta como para tratarla en cursos completamente distintos. De hecho, frecuentistas y bayesianos aparecen

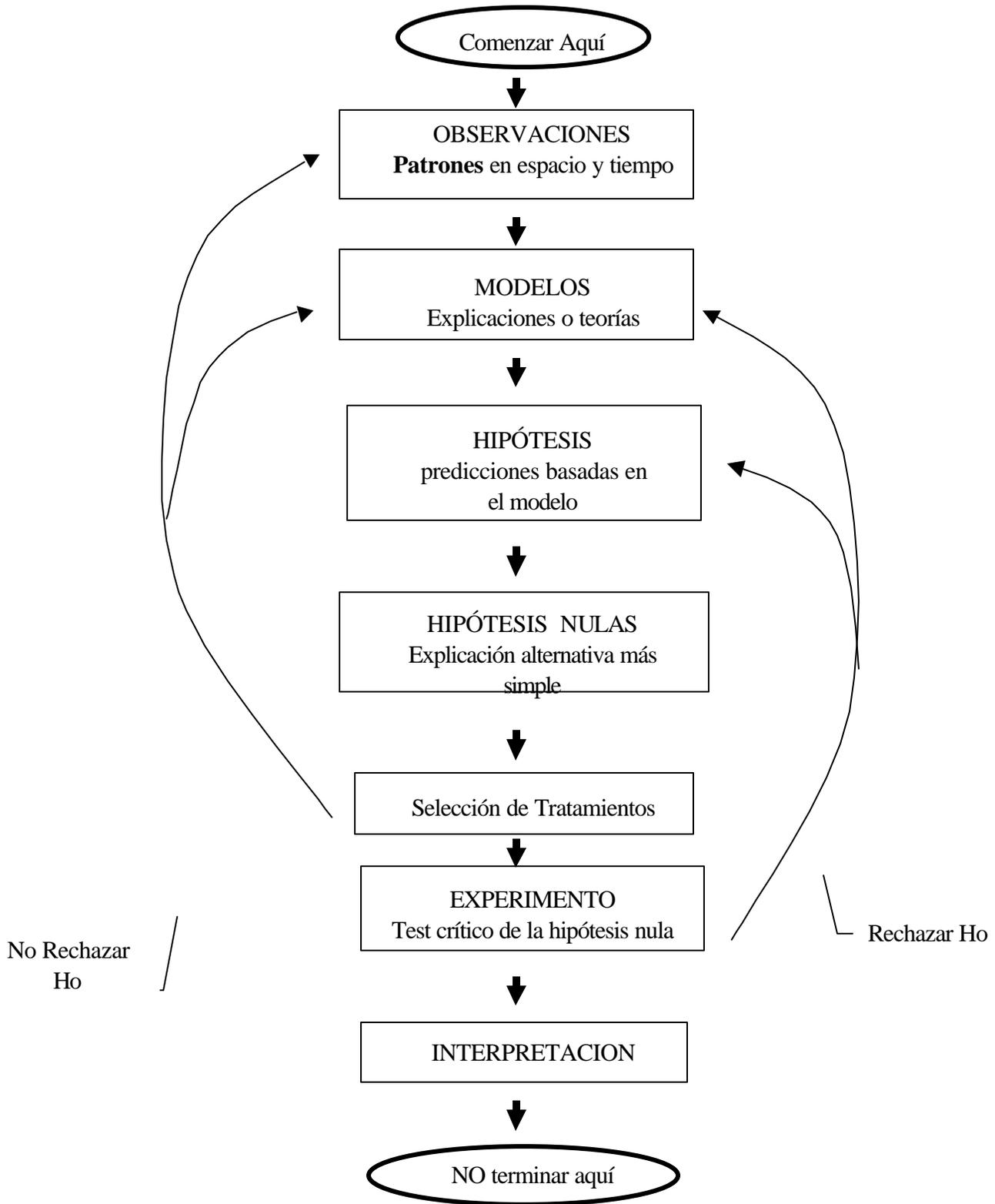
frecuentemente en lados opuestos de un campo de batalla que por momentos ha perdido incluso el tradicional decoro académico. En la última clase del curso trataremos de presentarles una introducción a la aproximación Bayesiana y su potencial aplicación al problema de evaluación de impacto ambiental.

El principal fundamento de este curso es que el desarrollo de experimentos comparativos exitosos necesita de objetivos de investigación claramente definidos, los que pueden ser resueltos a través de la elección apropiada de un *diseño de tratamientos*. El *diseño experimental* debe enmarcarse claramente en un diseño de investigación desarrollada por el o los investigadores y debe acomodar el diseño de tratamientos más eficiente en el contexto de las preguntas de esos investigadores, tanto desde el punto de vista estadístico como logístico.

En el proceso científico los investigadores deben seguir un *programa de investigación* claramente definido. Este programa de investigación se centra en el estudio de determinados “patrones”, que pueden ser la distribución de depredadores y presa en el bosque Maulino o la estructura de proteínas en células cancerosas. El programa de investigación debe generar las preguntas e hipótesis que se desea someter a prueba en cada etapa del programa. Estas hipótesis determinarán la elección del diseño de tratamientos, los que efectivamente apunten a responder la hipótesis en cuestión. Finalmente, los investigadores deben seleccionar el diseño experimental que incluye a los tratamientos y que facilita la recolección de los datos necesarios para evaluar la hipótesis, a la vez que permite controlar por las fuentes de error experimental y mantener constante la tasa de error Tipo I.

En este curso trataremos de hacer la conexión lo más clara posible entre las preguntas de los investigadores (objetivos del estudio), la selección de tratamientos (diseño de tratamientos) y la selección del mejor diseño experimental que nos permita controlar de mejor forma el error experimental. Luego de seleccionado el diseño experimental, en este curso veremos como podemos analizar los resultados usando las técnicas de Análisis de Varianza. Como resultado de este procedimiento, deberemos tomar la decisión de aceptar o rechazar una hipótesis nula que es el opuesto lógico a nuestra hipótesis de interés (hipótesis alternativa). Esta decisión nos llevará siempre a revisar nuestro modelo y/o a generar nuevas hipótesis que puedan explicar nuestras observaciones.

Aplicabilidad de estadística en un programa de investigación básica o aplicada:



Modificado de Underwood 1997

El proceso de investigación científica no es en absoluto tan lineal y organizado como representa este esquema. El esquema y direccionalidad puede entenderse como el resultado “neto” del proceso, luego de repeticiones y refinamientos entre todos los niveles.

Antes de empezar a tratar las materias del curso, es bueno mencionar una de las limitaciones de los cursos lectivos. El análisis de datos debe ser un proceso *interactivo*. Uno no puede seguir los pasos del libro de recetas a ciegas y llegar al plato final. Esto tampoco significa que uno debe “amasar” los datos de manera de alterar las conclusiones. El proceso interactivo significa que uno debe familiarizarse con los datos de un experimento o estudio, conocer bien la distribución de los datos, los problemas con algunas observaciones, como esos problemas pueden afectar nuestras conclusiones, etc. Es difícil resaltar suficientemente la importancia que tiene el graficar datos. Hacer buenos y muchos gráficos (no solamente los gráficos finales del informe o publicación) es una parte esencial del proceso de análisis e interpretación de resultados de cualquier estudio. Hoy en día, con los muchos paquetes computacionales que permiten hacer gráficos en forma muy rápida, no existe ninguna excusa para no explorar gráficamente los datos de un estudio. Lamentablemente, es difícil hacer gráficos exploratorios en las clases y por ello su importancia puede aparecer disminuida en las clases lectivas. Durante las sesiones prácticas trataremos de suplir esta deficiencia.

En esta clase repasaremos algunas definiciones y conceptos que ustedes deben asimilar para poder entender lo que viene más adelante.

DEFINICIONES

Población Estadística:

Es la totalidad de observaciones independientes (mediciones de la variable en que estamos interesados), acerca de la cual deseamos hacer inferencias (decir algo acerca de sus propiedades).

Una población estadística existe en un espacio y tiempo determinado y es una entidad real, pero sus límites pueden ser inconmensurables (ej. todo el universo). Los límites de una población estadística quedan definidos por nuestro interés o propósito.

Observaciones:

Las observaciones son nuestros “datos”: Medidas tomadas en la menor unidad de muestreo posible. Esta unidad es frecuentemente, pero no siempre individuos biológicos. La unidad de observación puede no ser equivalente a la *unidad experimental*, por lo que no siempre las observaciones son las réplicas del estudio.

La propiedad medida en cada unidad mínima de muestreo es nuestra **variable, carácter, o atributo**.

Ejemplos:

- Altura de estudiantes de la PUC
- Edad de egreso de estudiantes de doctorado
- Sueldo de profesores universitarios en Chile
- Una de las tres mediciones en un espectrofotómetro

Una población estadística está constituida por la totalidad de las observaciones de interés para los investigadores, las que pueden o no ser todas las observaciones de ese tipo existentes en el universo.

Una población estadística tiene ciertas características propias que la describen. Estas características de la población total se llaman **parámetros**.

Parámetros:

Propiedades o características fijas y únicas de una población estadística (e.g. media, varianza, etc.). Los parámetros son inamovibles y describen una población.

¿Cómo podemos conocer los parámetros de una población?

La única manera de “conocer” con certeza (100 % seguridad) los parámetros de una población a través de un censo, es decir la medición de la variable bajo estudio en todas las unidades de muestreo (e.g. individuos) de la población.

Cuando No podemos realizar un censo, necesitamos realizar una **inferencia** o **estimación** acerca del valor de estos parámetros, a través de tomar un **muestra** de la población de interés.

Inferencia Estadística:

Inferencia estadística es el proceso de hacer enunciados o afirmaciones acerca de una **población** basados en los resultados obtenidos en las **muestras** de dicha población.

Estadísticos/ Estadígrafos:

Mientras los parámetros describen una población total, o son los valores “reales” de la población, aquellos obtenidos de una muestra de dicha población se llaman **“estadísticos” o estadígrafos**.

Entonces, inferencia es el proceso de usar “**estadígrafos o estadísticos**” obtenidos de una muestra (ej. media muestral, m) para estimar **parámetros** (ej. media poblacional, μ) de una población.

Ejemplo: una inferencia estadística sería el usar los datos de altura de los estudiantes de esta clase para **estimar** la altura promedio de estudiantes de biología de la Universidad Católica.

Estimadores Sesgados y NO sesgados:

Un estadístico no sesgado nos entrega una buena estimación de una parámetro: la mitad del tiempo estará por sobre el valor real y la otra mitad por debajo del valor real del parámetro.

Un estadístico no sesgado se aproximará al valor real del parámetro a medida que aumentamos el tamaño muestral (n).

Un estadístico sesgado entrega siempre un valor pro debajo o por arriba del valor del parámetro real.

Aquí usaremos letras griegas para representar parámetros y letras latinas para representar estadísticos.

Replicación

Replicación básicamente significa la repetición independiente del mismo experimento básico. Más específicamente, cada tratamiento o condición es aplicado independientemente a cada uno de dos o más unidades experimentales.

Así, una *unidad experimental* es la mínima unidad de observación que puede considerarse como “independiente” de otras observaciones. La varianza de las unidades experimentales es Varianza del Error Experimental. Muchas veces se confunde esta varianza con la varianza de observaciones múltiples tomadas en la misma unidad experimental. La varianza de las observaciones dentro de las unidades experimentales no provee una estimación del error experimental.

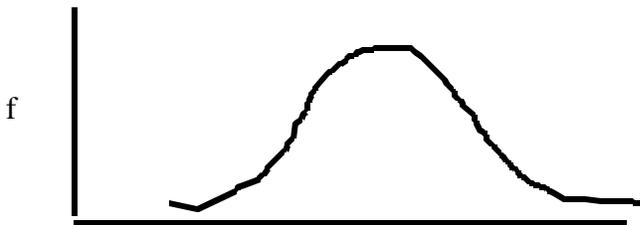
Existen varias razones para replicar un experimento:

- La replicación demuestra que los resultados son reproducibles, al menos bajo las condiciones experimentales existentes

- La replicación provee un grado de seguro frente a resultados aberrantes en el experimento debido a factores no previstos (factores “demonicos” de Hurlbert 1984)
- La replicación provee una manera de estimar la varianza del error experimental. Aunque experimentos previos hayan estimado esta varianza, la varianza del presente experimento puede ser más exacta pues refleja las condiciones del experimento
- La replicación provee la capacidad de aumentar la precisión de nuestras estimaciones de las medias de los tratamientos (estimación de parámetros). Al aumentar la replicación r veces disminuye el error estándar de Y en σ^2/r .

DESCRIPTORES DE UNA POBLACIÓN Y MUESTRA:

Existen dos maneras básicas de describir una población estadística:



A. Medidas de **localización** o ubicación, también llamadas **medidas de tendencia central**.

Describen la posición de nuestra muestra o población a lo largo de alguna dimensión (variable).

Las medidas de tendencia central más usadas son:

1. **Media aritmética** (‘promedio’):

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^n y_i$$

y_i = observación i

\bar{Y} = Media muestral

μ = Media Poblacional

La media aritmética tiene una propiedad muy interesante: Tiende a estar normalmente distribuida aún cuando viene de poblaciones que no están normalmente distribuidas.

La media es sensible a la presencia de valores extremos o “outliers”

\bar{Y} es un estimador NO sesgado de μ

2. Media Geométrica:

$$GM = \text{anti log} \frac{1}{n} \sum_{i=1}^n \log y_i \longrightarrow GM = \sqrt[n]{\prod_{i=1}^n y_i}$$

3. Media Armónica:

4. Mediana:

Valor de la variable que tiene igual número de observaciones a la izquierda (valores inferiores) y a la derecha (valores superiores). Divide una distribución de frecuencia en dos mitades con igual número de observaciones.

Es menos sensible que la media a la presencia de valores extremos.

5. Moda:

Valor(es) que más se repite en una distribución: el valor más popular.

B. Medidas de Dispersión:

1. Rango:

Valores extremos de una variable. Extremadamente sensible a “outliers”

2. Varianza:

$$s^2 = \frac{\sum_{i=1}^n (y_i - \bar{Y})^2}{n}$$

Desviaciones (deviates): $y_i - \bar{Y}$ Asumirá valores positivos y negativos (y)

Necesitamos sumar todas las desviaciones.

⇒ ¿Qué pasa si sumamos las desviaciones de la media? Deben dar como resultado cero.

⇒ Necesitamos sumar los cuadrados de las desviaciones.

⇒ ¿Qué pasa si usamos otra referencia que no sea la media para ver dispersión?

⇒ La suma de cuadrados será siempre mayor

3. Desviación Estándar

$$s = \sqrt{\text{varianza}} \longrightarrow \sqrt{s^2}$$

Puesto que para calcular la varianza debemos elevar las desviaciones al cuadrado, las desviaciones estarían expresadas en unidades de la variable de interés al cuadrado. La desviación estándar es una medida de dispersión expresada en las mismas unidades de la variable en cuestión.

Se ha demostrado que $s = \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{Y})^2}{n}}$, la desviación estándar muestral, es un estimador **SESGADO** de la verdadera desviación estándar poblacional (sigma). Entonces, se puede demostrar que:

$$s = \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{Y})^2}{(n-1)}} \quad \text{Es un estimador NO sesgado de la verdadera desviación estándar poblacional.}$$

La cantidad (n-1) se denomina **GRADOS DE LIBERTAD**.

Grados de Libertad:

Son una medida del nivel de confianza real que tenemos en la estimación de un parámetro poblacional. Con frecuencia, los verdaderos grados de libertad son inferiores al tamaño muestral, n.

Más adelante veremos una definición más formal de grados de libertad y como calcularlos.

4. Coeficiente de Variación

Puesto que la magnitud de la desviación estándar dependerá de a) las unidades de medida y b) de la posición o media de la variable (población de individuos más grandes tienen un mayor rango posible de variación), si queremos comparar el grado de variabilidad entre poblaciones podemos usar el **coeficiente de variación**.

$$CV = \frac{s \times 100}{\bar{Y}}$$

5. Error Estándar

Una medida normalmente confundida con la desviación estándar.

El error estándar es la desviación estándar de un estadístico tomado de varias muestras de una misma población:

Por ejemplo: El error estándar de la media es la desviación estándar de una serie de medias muestrales de una población determinada. A cada muestra independiente se debe calcular la media y de estas medias se calcula la desviación de la gran media.

Lo anterior es equivalente a:

$$s_{\bar{Y}} = \frac{S}{\sqrt{n}} \quad \text{y su correspondiente estimador muestral} \quad SE = \frac{s}{\sqrt{n}}$$

El error estándar es una función de a) la desviación estándar de las observaciones, así como del tamaño de la muestra. Al aumentar el tamaño muestral, disminuimos el error estándar.

MUESTREO ALEATORIO Y DISTRIBUCION DE PROBABILIDADES

La *teoría de probabilidad* es la base de toda la ciencia estadística, tanto para la prueba de hipótesis como para la estimación o inferencia.

Aquí no vamos a ver nada de teoría de probabilidades, si ustedes no se acuerdan tendrán que repasar sus textos básicos. Es responsabilidad de ustedes entender que son las combinaciones y permutaciones.

Distribución de Probabilidad:

Una distribución de probabilidad nos indica la “**probabilidad**” de observar un **evento** determinado, perteneciente a un **espacio muestral** definido y dado un **tamaño muestral** determinado.

Por ejemplo, al lanzar una moneda, los eventos posibles son cara y sello y esto define el espacio muestral. La probabilidad de que una serie de 100 lanzamientos (tamaño muestral) de la moneda 50 sean cara y cincuenta sean sello (evento: 50 caras y 50 sello) esta definida por una distribución de probabilidad que nos determina cuantas veces podemos encontrar este evento **por simple azar**.

Requisito para usar cualquier distribución de probabilidad: Tener una representación FIEL de la población (o espacio muestral). Por ejemplo, NO podemos usar una moneda que tenga un sesgo hacia cara o sello. debemos usar un muestreo NO sesgado y la única manera de asegurar que tendremos una representación fiel de la población es a través de un **muestreo aleatorio**.

Muestreo Aleatorio:

Representación FIEL de una población en un subset de observaciones. Siempre involucra muestreo aleatorio (asignación aleatoria de tratamientos, cuadrantes, etc.) a algún nivel. Lo importante es asegurar que cada observación individual que forma parte de la población tenga IGUAL oportunidad o probabilidad de ser incorporada en la muestra.

La muestra NO puede ser sesgada hacia un sector de la población. Falta de muestreo aleatorio NO tiene solución en ningún análisis estadístico.

DIFERENTES TIPOS DE ESTUDIOS: *OBSERVACIONALES* Y *EXPERIMENTALES*

Existe una distinción importante entre los tipos de estudios o investigaciones que uno puede realizar, esto es entre los estudios “observacionales” y los estudios “experimentales”. No siempre hay una línea divisoria clara y precisa entre estos tipos de estudios, pero es importante mantener en mente sus diferencias.

En los estudios *observacionales*, o más propiamente *comparativos observacionales* los datos u observaciones son recolectadas a través de la observación de un proceso que puede no ser bien entendido. En general, la existencia del proceso, así como su efecto sobre la variable respuesta de interés, son ambos “evidenciados” por las observaciones que se recolectan.

Por ejemplo, los registros de hospitales pueden ser estudiados para ver si la incidencia de una enfermedad esta relacionada o no a la presión arterial. Uno puede estudiar la correlación entre un depredador y su presa a través del espacio para inferir el potencial efecto del depredador sobre la presa.

Los estudios experimentales involucran por lo general la recolección de datos sobre un procesos cuando hay alguna manipulación de variables que se supone que afectan el resultado del proceso, manteniendo otras variables constantes.

Por ejemplo, asignando pacientes a diferentes drogas y observando sus tiempos de reacción o manipulando la presencia de depredadores.

En muchos casos las mismas técnicas estadísticas pueden usarse para analizar datos observacionales y experimentales. Sin embargo, la validez de las inferencias que resultan del análisis dependerá de la naturaleza de los datos.

Un efecto que es observado en forma consistente en las replicas de un experimento manipulativo puede ser explicado razonablemente por la manipulación de la variable experimental o tratamiento.

Por ejemplo, si la mortalidad de ratones es siempre mayor en las 10 réplicas (cajas), seleccionadas al azar, en que alimentamos ratoncitos con “chépica” en comparación a otras 10 replicas bajo una dieta sin chépica, es razonable suponer que la chépica nos esta matando los ratones.

Por otro lado, en un estudio observacional la misma consistencia de los datos puede producirse porque todos los datos son afectados de la misma manera por una variable desconocida y no medida.

Por ejemplo, si comparamos la sobrevivencia de ratones en 10 laderas de cerros con chépica y 10 laderas sin chépica, y encontramos que la sobrevivencia de ratones es mayor en laderas sin chépica, la conclusión es solamente “buena evidencia” de que la chépica os esta matando los ratones. Es muy razonable también suponer que cualquiera sea el factor que determina la abundancia de chépica en primer lugar, también afecta a los ratones.

De los muchos problemas potenciales que pueden afectar a estudios observacionales, tal vez el más importante es que es difícil obtener una conclusión limpia y directa por el efecto “*confounding*” (confundidor) de las variables “no controladas”. Es muy difícil saber si los efectos observados en los datos son el producto de cambios en la variable de interés o a cambio que también ocurren en otras variables.

Otro problema frecuente con la estimación realizada a partir de estudios observacionales se produce por la dificultad de realizar un muestreo no sesgado y realmente aleatorio. Es muy fácil confundir muestreo “conveniente” con completamente aleatorio.

Un tercer problema frecuente ocurre cuando las observaciones tienen que realizarse sobre grupos de individuos, en lugar de sobre los individuos mismos. Estudios experimentales no están necesariamente exentos de este problema, pero es menos frecuente debido a la naturaleza de los experimentos. Por ejemplo, un estudio puede analizar la mortalidad de ratones en varios sitios, en relación al consumo promedio de chépica y encontrar que la sobrevivencia de los ratones es menor en lugares con más chépica. La conclusión es que el consumo de chépica aumenta la mortalidad. Sin embargo, esta no es la única conclusión posible ya que es fácil demostrar que las relaciones que aplican a los individuos dentro de grupos pueden no aplicar a las relaciones entre grupos. Este cambio en las relaciones entre variables a nivel de individuos versus grupos de individuos se denomina “falacias ecológicas”. Por ejemplo, es posible que un sitio presente una alta mortalidad y un alto consumo de chépica, pero que los ratones que se mueren no son los que consumen chépica.

De lo anterior uno puede concluir que el mejor entendimiento estaría basado en experimentos y no en simples observaciones. Sin embargo, los estudios también tienen muchas

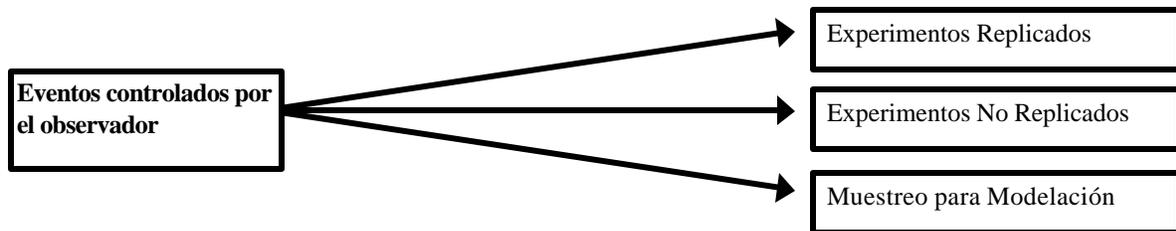
restricciones. Tal vez la más importante, es la dificultad de capturar la complejidad del mundo real en un experimento que necesariamente debe ser simplificado.

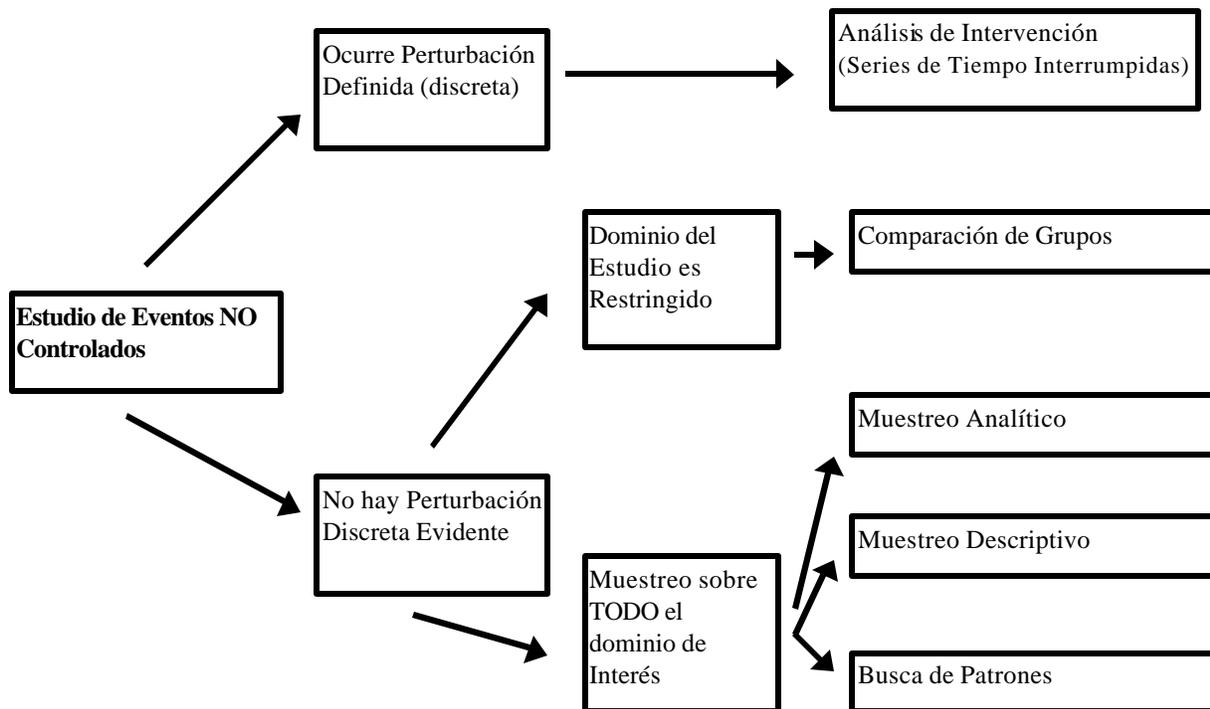
Por ejemplo, uno puede demostrar que ratones sometidos a una dieta de pura chéptica sobreviven menos que ratones bajo una dieta alternativa. Pero si la chéptica no es más que uno de los 57 tipos de plantas y 32 tipos de insectos que los ratones consumen regularmente en el campo, ¿Nos sirve este resultado para explicar las diferencias en sobrevivencia de ratones entre poblaciones naturales?

Un experimento debe entonces ser diseñado de manera muy cuidadosa y tratando de establecer un balance entre el grado de “control” que se quiere establecer sobre las variables “no manipuladas” (temperatura, fotoperiodo, etc.) y el grado de realismo del experimento que permita la correcta interpretación de las consecuencias de la variable manipulada.

Otra complicación es simplemente las limitaciones logísticas o éticas para desarrollar estudios experimentales. En ecología, por ejemplo, una enorme limitación esta dada por la escala espacial y temporal en la cual se pueden desarrollar experimentos, así como por el grado de complejidad que es posible capturar en un estudio experimental determinado.

Aunque yo no soy muy amigo de las clasificaciones porque estas tienden a tomarse de forma rígida, es posible clasificar subdividir los diferentes tipos de estudios de una manera didáctica, siguiendo la clasificación de Eberhart & Thomas (1991):





La separación entre eventos controlados por el Observador o No controlados por el observador es equivalente a la separación entre estudios experimentales y observacionales. Los modelos que veremos en este curso han sido desarrollados principalmente para casos en que el observador puede manipular las variables y puede realizar estudios replicados. En algunos casos, aún y cuando se desarrollen estudios o manipulaciones experimentales, no es posible replicar la aplicación de tratamientos y esto da origen a una serie de técnicas estadísticas muy distintas a las más tradicionales. En algunas otras circunstancias, estudios experimentales son desarrollados con el solo fin de parametrizar un modelo y allí el énfasis es puesto en estimación de parámetros de un experimento en lugar de prueba de hipótesis.

En muchas circunstancias, uno puede claramente distinguir una perturbación discreta y distintiva en un estudio observacional, aún y cuando la perturbación no haya sido realizada por el observador. En estos casos se puede usar técnicas estadísticas denominadas ‘análisis de intervención’.

Cuando no hay una perturbación discreta evidente existen varias alternativas dependiendo de si uno cuenta o no con toda la población de interés. Uno de los estudios más comunes es cuando uno tiene una “muestra” de casos u observaciones con y sin la variable de interés y puede así realizar comparaciones entre grupos. En estos casos es extremadamente importante que las observaciones usadas para someter a prueba la hipótesis de interés NO sean las mismas que dieron origen al estudio en primer lugar!

Los experimentos “de verdad” no son fáciles de desarrollar y la diferencia entre los diferentes tipos de estudio de Eberhart & Thomas es mucho más borrosa de lo que aparece en un esquema. En la mayoría de las circunstancias estamos obligados a realizar *quasi-experimentos*, los que a pesar de ser más limitados que los estudios experimentales de verdad, tienen ventajas sobre los estudios meramente observacionales.

Por ejemplo, si queremos estudiar el efecto de fumar tabaco sobre la frecuencia de enfermedades cardíacas, el estudio ideal involucra seleccionar individuos al azar de la población de interés (ej. Chilenos adultos) y luego asignar al azar la mitad de los individuos a “fumadores” y la otra mitad a “no fumadores”, y luego observar la frecuencia de enfermedades cardíacas. Obviamente, desarrollar este experimento puede ser poco ético pues implica obligar a fumar a la mitad de los individuos. Una alternativa a este estudio es simplemente analizar la frecuencia de enfermedades cardíacas en grupos de fumadores y no fumadores, desarrollando así un estudio observacional. El problema es que pueden existir (y de hecho existen) muchas variables asociadas con el hecho de fumar, las que pueden tener asociadas dolencias cardíacas, no directamente relacionadas al cigarrillo. Un compromiso entre estas aproximaciones puede ser el persuadir algunos fumadores que dejen de fumar y comparar sus subsecuentes enfermedades cardíacas en comparación a los que no dejaron de fumar. El problema potencial aquí está dado por la predisposición a dejar de fumar, la cual puede estar correlacionada con las enfermedades cardíacas y nos produce un potencial “*confounding*” de la variable de interés (fumar o no fumar). Esto se produce por la imposibilidad de asignar en forma completamente aleatoria los tratamientos a las unidades experimentales.

¿QUÉ TIPO DE PREGUNTAS CIENTÍFICAS SE RESPONDEN A TRAVÉS DE LA PRUEBA DE HIPÓTESIS Y LOS INTERVALOS DE CONFIANZA?

Existen dos preguntas fundamentales que los investigadores deben responder durante el curso de una investigación científica:

A. ¿Cuán *confiables* son los resultados obtenidos?

Por ejemplo, si durante un experimento uno aplica una hormona de crecimiento a un grupo de chanchos y éstos crecen 10 kilos más al cabo de un mes que aquellos chanchos sin hormonas de crecimiento, ¿Cuán confiable es que el efecto *real* de la hormona de crecimiento sea efectivamente 10 kilos?

En otras palabras, si aplicamos esta hormona a todos los chanchos del universo y de esa manera *conocemos* el efecto real (parámetro) de la hormona en un censo de los chanchos, ¿Cuán acertada sería nuestra estimación de 10 kilos al mes?

Para responder este tipo de preguntas nosotros usamos **Intervalos de Confianza** y hacemos *inferencias estadísticas*.

Intervalo de Confianza:

Intervalos de confianza son una medida de la certidumbre (confiabilidad) que nuestro estadístico se aproxime al valor *real* poblacional. Los intervalos de confianza expresan la *probabilidad* que los *límites* definidos por el intervalo incluyan efectivamente el valor real (parámetro).

Por ejemplo, si nuestra estimación del efecto de la hormona de crecimiento es 10 kilos y con nuestros datos calculamos que el *intervalo de confianza* al 95% es 3.5, entonces podemos decir que: “existe un 95% de probabilidad de que el intervalo entre 6.5 y 13.5 kilos efectivamente contenga la media real de la población”

B. La otra pregunta fundamental durante el curso de una investigación científica, particularmente en investigación básica más que aplicada es:

¿Cuán probable es que las diferencias entre los resultados observados y esperados bajo la base de una hipótesis particular hayan sido *producidos por simple azar*?

En el ejemplo anterior, Cuán probable es que por simple azar el grupo de chanchos que recibió la hormona de crecimiento haya crecido 10 kilos más que el grupo control en un mes?

Este tipo de preguntas acerca de la “significancia” estadística (valor de P) de un resultado se responde a través de la **Prueba de Hipótesis**.

Prueba de Hipótesis:

Una prueba de hipótesis estadística es tomar la decisión de aceptar o rechazar una *hipótesis nula*, cuantificando la *probabilidad de cometer un error* al tomar esta decisión y usando un criterio *arbitrario pre establecido*.

Por ejemplo, si seguimos el standard de considerar *significativo* algo que por simple azar no ocurre más de 1 en 20 veces (5% de las veces), entonces tomamos la decisión de rechazar una hipótesis nula (que las diferencias entre grupos de chanchos no son significativas) cuando nuestra probabilidad de error es menor del 5% de las veces.

El ámbito de la aplicación estadística en la Toma de Decisiones.

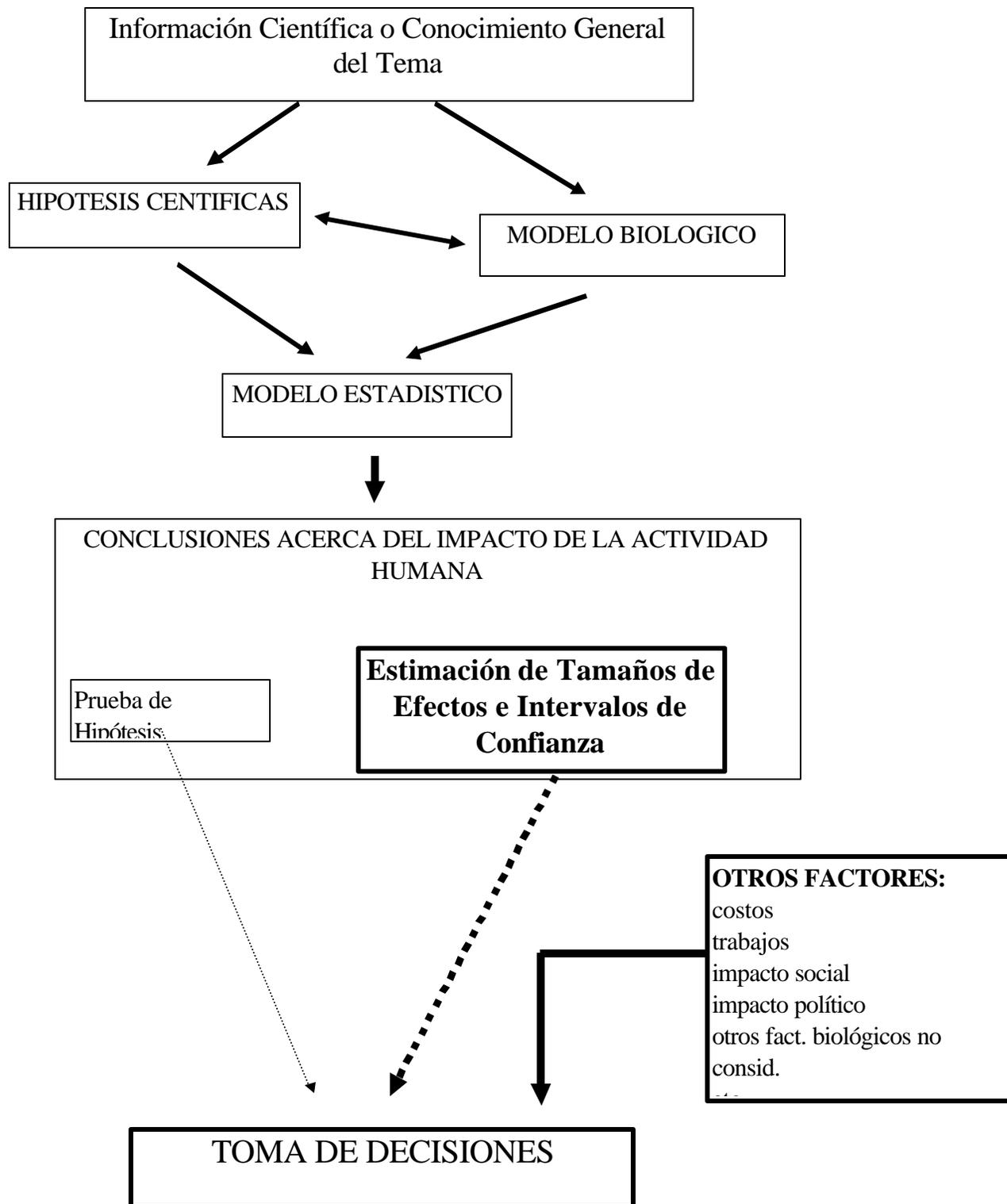
Hasta ahora cuando nos referimos a “tomar una decisión” nos hemos referido solamente a la decisión de aceptar o rechazar una hipótesis estadística determinada, la cual tiene un fundamento o modelo biológico.

Esta decisión de aceptar o rechazar una hipótesis basados en un nivel preestablecido arbitrariamente (5, 10%, 20%) **NO** es lo mismo que la “Toma de Decisiones” en el proceso de evaluación de impacto ambiental o en general en la toma de decisiones para la regulación de una actividad humana.

Esto parece bastante trivial, sin embargo, mucha gente no especialista y también algunos pseudo especialistas en evaluación ambiental confunden estas dos cosas.

En general la *toma de decisiones* es un proceso mucho más complejo. En esta toma de decisiones acerca de una actividad humana concurren muchos aspectos biológicos y no biológicos y ciertamente científicos y no científicos. La prueba de hipótesis estadística tiene (o debería tener) un rol muy reducido en este proceso, pero se emplea con mucha (demasiada) frecuencia en evaluación de impacto ambiental. La “inferencia” estadística de tamaños de efectos (incluyendo intervalos de confianza) provee mayor información en el proceso de toma de decisiones.

IMPACTO DE ACTIVIDADES HUMANAS



TEXTOS SUGERIDOS:

En este curso no existe un texto guía. En primer lugar, existen pocos libros de texto que tengan el nivel de profundidad y de cobertura apropiados para un curso introductorio de diseño experimental para biólogos. En segundo lugar, los pocos libros que sí se han escrito con esta intención no están aún disponibles en el idioma español. Habiendo dicho esto, hay varios libros que de consulta que pueden ser muy útiles para profundizar más las materias tratadas en esta clase o aclarar conceptos. De estos yo recomiendo:

Kuehl, R. O. 1994. Statistical principles of research design and analysis. Duxbury Press, Belmont, California.

Muy buen texto, pero algo más avanzado que lo que veremos en este curso.

Manly, B. F. J. 1992. The design and analysis of research studies. Cambridge University Press, Cambridge.

Buenos ejemplos de casos de estudio. No muy bueno como texto.

Mead, R. 1988. The Design of Experiments: Statistical Principles for Practical Applications. Cambridge University Press, Cambridge.

Muy buen tratamiento de diseño experimental en bloques.

Sokal, R. R., and F. J. Rohlf. 1981. Biometry, Second Edition. W.H. Freeman & Co., New York.

Excelente libro de referencia. Hay versiones más nuevas y también en español.

Sprenst, P. 1993. Applied nonparametric statistical methods, 2nd Edition. Chapman & Hall, London.

Libro general para métodos de distribución libre

Underwood, A. J. 1997. Experiments in ecology. Cambridge University Press, Melbourne, Australia.

Buen texto guía, pero con mucho de la visión muy propia del autor.

Williams, B. 1993. Biostatistics. Concepts and applications for biologists. Chapman & Hall, London.

Descripciones claras de conceptos básicos, incluyendo distribuciones de probabilidades.

Clase 2: Clases de Análisis Estadísticos, Intervalos de Confianza, Prueba de Hipótesis, Error Tipo I y Tipo II

1. CLASES DE ANÁLISIS ESTADÍSTICOS

Hoy en día existe un gran número de técnicas estadísticas que son usadas con mayor o menor frecuencia en estudios científicos y aplicados. Algunas de ellas han sido desarrolladas para responder preguntas sumamente específicas dentro de un área de las ciencias biológicas particular. Otras son de aplicación más general.

Resulta muy fácil ahogarse en este mar de técnicas y aproximaciones estadísticas al análisis de datos, aun para los ya iniciados en el tema. Frente a este océano, a veces es útil tener una guía general que ayude a despejar algunas de las opciones. Con este animo, aquí les presento una clasificación general de métodos estadísticos que ha sido modificada de McCune 1992, Sokal & Rohlf 1981 y Manly 1998. Esta clasificación dicotómica es muy incompleta y menciona solamente las técnicas más usadas o que me son más familiares.

Clave Dicotómica para ayudar a identificar la Clase de Análisis Estadístico Apropriado

1. Una Muestra (no hay partición en grupos de unidades de muestreo)

1.A. Sin separación entre variables independientes y dependientes

1.A.1. Sin variables categóricas

1.A.1.a. Dos variables

**Correlación de Pearson
(Correlación Kendall, Spearman)**

1.A.1.b. Más de dos variables

1.A.1.b.1. Un set de variables

**PCA, CA, DCA, CA
(Bray-Curtis, NMDS)**

1.A.1.b.2. Dos sets de variables

**Correlación Canónica
Análisis de Correspondencia Canónica
Análisis de Redundancia**

1.A.2. Alguna o todas la variables Categóricas

1.A.2.a. Todas las variables categóricas

1.A.2.a.1. Dos variables

**Tablas Contingencia de Dos vías
(Chi-cuadrado, G-test)**

1.A.2.a.2. Más de Dos variables

**Tablas de Contingencia multi-vías
(G-test, Chi-cuadrado)**

1.A.2.b. Algunas Variables categóricas y otras ordinales

Usar variables categóricas para definir grupos
Ningún análisis completamente apropiado

1.B. Variables separadas en Dependientes e Independientes

1.B.1. Una Variable Dependiente

1.B.1.a. Variable Dependiente Categórica

1.B.1.a.1. Una Variable Independiente

Regresión Logística Simple

1.B.1.a.2. Más de una Variable Independiente

Regresión Logística Múltiple

1.B.1.b. Variable Dependiente Ordinal

1.B.1.a.1. Una Variable Independiente

**Regresión Lineal Simple
Regresión Estructural
Regresión 'Reduced Major Axis'**

1.B.1.a.2. Más de una Variable Independiente

Regresión Múltiple

1.B.2. Más de una Variable Dependiente

Regresión Canónica

2. Dos o más muestras (muestra particionada en grupos de unidades de muestreo)

2.A. Una Variable Dependiente Ordinal y Variables Independientes Categóricas

2.A.1. Dos Muestras

2.A.1.a. Muestras Pareadas

**Prueba t-Student Pareada
ANDEVA con bloques
(Prueba Wilcoxon signed rank)**

2.A.1.b. Muestras No Pareadas

**Prueba t-Student
ANDEVA una vía
Kolmogorov-Smirnov para dos muestras
(U-Mann & Whitney)**

2.A.2. Más de dos Muestras

2.A.2.a. Clasificación Simple (Completamente Aleatorio)

**ANDEVA una vía
(STP)
G-test para frecuencias
(Kruskal-Wallis)**

2.A.2.b. Clasificación Anidada

ANDEVA anidado

2.A.2.c. Clasificación Factorial

ANDEVA factorial

2.A.2.d. Clasificación en grupos de unidades similares

ANDEVA con bloques

2.B. Una Variable Dependiente Ordinal, Variables Independiente incluye Categórica (clasificación) y Ordinal

ANCOVA

2.C. Dos o más Variables Dependientes

MANOVA

2.D. Sin separación entre dependientes e independientes

2.D.1. Clasificación de unidades muestrales en nuevos grupos

**Análisis de Cluster jerárquico y no jerárquico
Twinspan**

2.D.2. Evaluación del poder discriminatorio de los grupos

**Análisis Discriminante
Multiple Response Permutation Procedure (MRPP)**

2.D.3. Ordenación de muestras y grupos para reducir dimensionalidad y explorar patrones

PCA, DCA, CA, Bray-Curtis, NMDS

2. INTERVALOS DE CONFIANZA

Los intervalos de confianza nos permiten calcular el grado de certidumbre de nuestras estimaciones de determinados parámetros. El cálculo mismo de intervalos de confianza no implica someter a prueba hipótesis acerca de las diferencias entre grupos. No deben ser usados para responder o someter a prueba determinadas hipótesis, sino más bien para determinar la *magnitud* real de un parámetro. Esta magnitud puede ser el tamaño del “efecto” de un determinado tratamiento en un experimento, pero también puede ser el valor de un parámetro usado en un modelo de simulación.

El cálculo paramétrico de intervalos de confianza se basa en las propiedades de la distribución normal (revisar apuntes BIO-242A). Si queremos calcular el intervalo de confianza al 95 % alrededor de la media de una población, a partir de una muestra tomada en forma aleatoria de esa población, podemos usar la expresión:

$$\bar{Y} - 1.96s_{\bar{Y}} \leq m \leq \bar{Y} + 1.96s_{\bar{Y}}$$

La constante 1.69 esta dada por el número de unidades de desviación estándar bajo una curva normal estandarizada, bajo la cual se encuentra el 95% de las observaciones.

Esta expresión tiene la inconveniencia que requiere conocer el error estándar poblacional, lo que no es frecuente en la practica. Casi siempre estaremos enfrentados a situaciones en que no conocemos la media paramétrica de una determinada variable ni tampoco la varianza paramétrica. Por ello, debemos usar una estimación del error estándar basado en el error estándar de la muestra y además debemos corregir por desviaciones de normalidad causadas por pequeños tamaños muestrales. La expresión que usamos es:

$$\bar{Y} - EE * t_{[0.05, n-1]} \leq m \leq \bar{Y} + EE * t_{[0.05, n-1]}$$

En esta expresión, EE es el error estándar muestral y t es el valor de la distribución de t-Student evaluada a 1- % de Confianza, en este caso 1-0.95 = 0.05 y con los grados de libertad con se calcula la desviación estándar de la muestra (n-1).

Esta es una expresión general que sirve para calcular intervalos de confianza con distintos niveles de probabilidad de cualquier parámetro y con diferentes tamaños muestrales, siempre que se pueda suponer que la variable de interés sigue una distribución aproximadamente normal. El aspecto crítico para el cálculo de intervalos de confianza es obtener una estimación apropiada del Error Estándar del estadístico que se calcula en la muestra. Esto puede parecer trivial en los diseños más simples, pero es menos obvio en diseños más complejos y es muy común que investigadores usen el error estándar equivocado. Para calcular el error estándar en forma correcta se debe identificar cuales son las *unidades experimentales* (unidades independientes).

En este curso no veremos mucho más que esto acerca de intervalos de confianza ya que no tendremos tiempo para cubrir diseño experimental y pruebas de hipótesis. Pueden encontrar más información en Sokal & Rohlf 1981, Mead 1988, Williams 1996.

3. INTRODUCCIÓN A PRUEBA DE HIPOTESIS

La prueba de hipótesis es una de las aplicaciones más importantes y frecuente de la teoría estadística al desarrollo de las ciencias básicas y aplicadas y en particular en las ciencias biológicas y ambientales.

La aplicación de modelos estadísticos de prueba de hipótesis ha estimulado la rigurosidad en el desarrollo e interpretación de experimentos científicos y en los protocolos de control de calidad de la industria. Este desarrollo estadístico también ha permitido diseñar experimentos más complejos que intentan responder preguntas más específicas o complejas y que permiten controlar de mejor manera el error experimental. El mejor control del error experimental a través del diseño de los experimentos ha permitido una mayor precisión no sólo en la prueba de hipótesis, sino que también en la estimación de parámetros.

Como vimos anteriormente, existe un sin número de métodos estadísticos de prueba de hipótesis, tanto “paramétricos” como los llamados “no-paramétricos”, pero la filosofía básica de todos estos es muy similar.

Prueba de Hipótesis:

Una prueba de hipótesis estadística es tomar la decisión de aceptar o rechazar una *hipótesis nula*, cuantificando la *probabilidad de cometer un error* al tomar esta decisión y usando un criterio *arbitrario y pre establecido*.

También podemos entender una prueba de hipótesis como el proceso de examinar una *muestra* de la población bajo la base de una *distribución de datos esperada* de acuerdo a una *hipótesis* particular, lo cual lleva a una decisión de aceptar la hipótesis subyacente o rechazarla y aceptar una hipótesis *alternativa*. Existe un máxima que dice que:

*La Ciencia funciona a través de **Rechazar** Hipótesis Nulas y NO a través de **Aceptar** hipótesis alternativas*

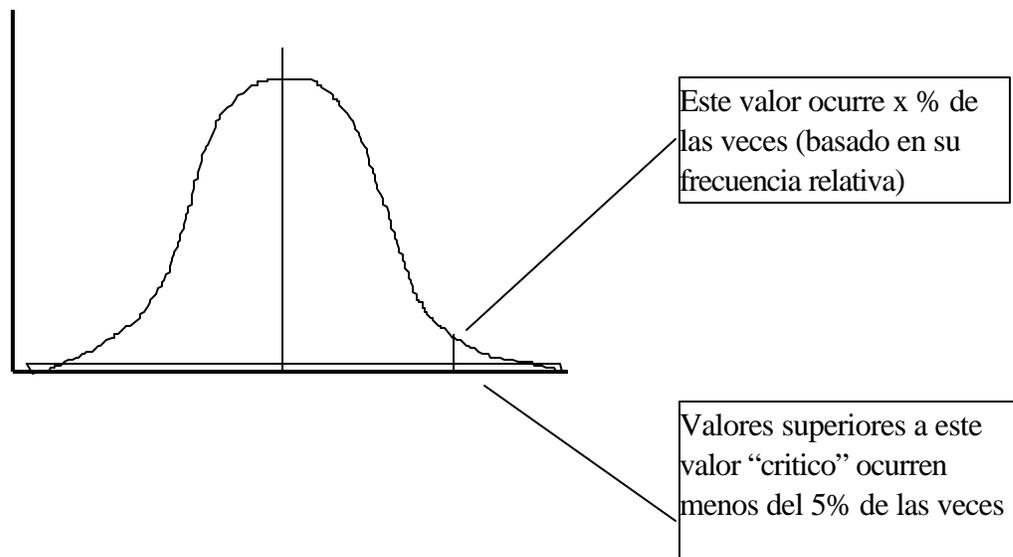
Puesto que normalmente la hipótesis en que estamos interesados, es decir aquella que nos entrega la mayor información biológica es la H_a , ¿Por qué no diseñamos estudios para *confirmar* esta hipótesis alternativa en ves de estudios para rechazar la hipótesis nula? Desarrollar un programa de investigación para corroborar hipótesis alternativas se ha denominado “Confirmación de Hipótesis” y a pesar de que ha sido en general fuertemente criticado por filósofos de las ciencias, la verdad es que las ciencias ecológicas y evolutivas no han estado ajenas a esta aproximación, particularmente en áreas que no permiten la manipulación experimental. En el curso BIO-242A presentamos algunas de las principales complicaciones con la metodología de confirmación de hipótesis. Los interesados pueden consultar los muchos libros y artículos de filosofía de las ciencias que han sido escritos en los últimos dos siglos. Para los menos inclinados a leer los áridos artículos filosóficos, hay un ensayo corto e interesante escrito por Paul Dayton titulado: “Ecology: A Science and a Religion too” .

4. PRUEBA DE HIPÓTESIS: Pruebas de Una y Dos Colas

Dijimos que una prueba de hipótesis es tomar una decisión acerca de la veracidad de una hipótesis nula, sabiendo cual es el margen o probabilidad de error al tomar dicha decisión.

Volvamos a nuestra distribución normal. En esta distribución, dijimos que podemos calcular exactamente las áreas por encima y por debajo de un valor determinado. Esto equivale a decir que podemos determinar la “probabilidad” que un determinado valor ocurra en esta población.

Figura 2.1.



Supongamos que nuestra experiencia nos sugiere que los estudiantes de biología son muy inteligentes y queremos saber si efectivamente los estudiantes de biología en universidades Chilenas tienen un CI más alto que el promedio de estudiantes universitarios.

Lo primero que tenemos que hacer es especificar las hipótesis en cuestión:

Ho: No existen diferencias entre el coeficiente intelectual los estudiantes de biología y el resto de los estudiantes universitarios

Ha: El CI de estudiantes de biología es más alto que el promedio de la “población” de estudiantes.

Esta pregunta equivale a preguntarse si los estudiantes de biología provienen o no de la misma población de estudiantes universitarios. De manera más formal, también podemos expresar nuestras hipótesis como:

Ho: $m = \mu$, donde m es la media de estudiantes de biología y μ la media poblacional de todos los estudiantes.

Si los biólogos fueran una muestra aleatoria, entonces el CI promedio debería ser cercano a 98.7, el promedio de todos los estudiantes universitarios (μ). Sin embargo, aún y cuando efectivamente los biólogos no sean más inteligentes que los otros estudiantes, el CI promedio no será nunca exactamente 98.7. La pregunta entonces es:

¿Cómo decidimos que tan diferente es aceptable? O, en otras palabras, ¿Cómo podemos decidir que las diferencias observadas no son producto de simple azar, como especifica la hipótesis nula?

Supongamos que el CI de los estudiantes universitarios sigue una distribución normal con media $\mu = 98.7$ y que los estudiantes de biología tienen una media $Y\text{-barra} (= m) = 115.4$.

Entonces podemos calcular la probabilidad de obtener este valor en la población. Sin embargo, en realidad no nos interesa la probabilidad de este valor o evento en particular, sino que la probabilidad de cualquier evento igual o mayor, pues todos los valores mayores rechazarán la hipótesis nula.

Es fácil entender por qué necesitamos calcular la probabilidad de todos los eventos entre un rango de valores. Por ejemplo, la probabilidad de que una familia de 12 hijos tenga 6 mujeres y 6 hombres es mucho menor de 50:50, de hecho es de sólo **22%**, la probabilidad de obtener 7 caras y 7 sellos en 14 lanzamientos de una moneda es de sólo **20%**. Esto ocurre porque estas son las probabilidades de eventos individuales y lo que queremos saber es cuál es la probabilidad de que el resultado de catorce lanzamientos este entre valores ‘aceptables’, digamos 4 : 10 y 10:4, este es el rango que consideramos “normal” o aceptable y de hecho lo denominamos el **rango de aceptación**.

Si calculamos la probabilidad de \bar{Y} igual o mayor a 115.4 y este valor resulta ser de $P=0.049$, ¿Qué podemos concluir?

Para poder tomar una decisión, DEBEMOS determinar un nivel de aceptación ANTES de realizar el estudio. Este es un nivel arbitrario que nos permite decir que un valor determinado es muy poco probable, por simple azar. Este valor crítico se denomina **Nivel de Significancia**.

En biología este criterio o **Nivel de Significancia** es casi universalmente fijado al 5%. Es decir, aquellos fenómenos que ocurren con una frecuencia menor al 5%, o una en 20 veces son considerados *significativos*.

El nivel de significancia determina nuestra zona de aceptación o rechazo. Es decir, si el valor observado es igual o superior al valor crítico, podemos concluir que los estudiantes de biología tienen un coeficiente intelectual más alto que otros estudiantes universitarios ¿Con qué probabilidad de error? Con un error de un 5% porque por simple azar esperamos que el 5% de la población de estudiantes universitarios tenga CI mayores al valor crítico.

¿Podemos decir que es *imposible* que estudiantes universitarios tengan valores tan altos como los de biología?...NO, solamente podemos decir que es poco probable que nuestra muestra de estudiantes de biología pertenezca a la misma población.

Las hipótesis anteriores pueden expresarse en forma estadística de la siguiente forma:

$$H_0: \mu_1 = \mu_2$$

$$H_a: \mu_1 > \mu_2$$

Esta hipótesis alternativa es de UNA COLA, y nuestra prueba de hipótesis es de una sola cola pues solamente estamos interesados en saber si los estudiantes de biología tienen CI más altos.

Si por el contrario nuestra observación sugiere que los estudiantes de biología son “distintos” a los estudiantes universitarios en general, pero NO sabemos si ellos tendrán CI más altos o más bajitos, entonces nuestra hipótesis correcta es:

$$H_0: \mu_1 = \mu_2$$

$$H_a: \mu_1 \neq \mu_2$$

Esta es una hipótesis y una prueba de hipótesis de DOS colas, por cuanto debemos mirar tanto a valores por encima como por debajo en la distribución de probabilidad.

Hipótesis Nula	Hipótesis Alternativa	Rango de Rechazo
$a = b$	$a > b$	valor crítico a ∞
$a = b$	$a < b$	$-\infty$ a valor crítico
$a = b$	$a \neq b$	$-\infty$ a valor crítico y valor crítico a ∞

5. TIPOS DE ERROR

Puesto que las respuestas biológicas son intrínsecamente variables, necesitamos hacer uso de probabilidades y estadística para tomar decisiones. Esto significa también que SIEMPRE existirá la posibilidad de cometer errores. La aplicación de estadística no significa que no cometeremos errores en tomar una decisión acerca de una hipótesis particular, sino solamente que podemos medir este error y “asumirlo”, a través de fijar un criterio de aceptación o nivel de significancia.

Como en todo orden de cosas en donde se toman decisiones bajo incertidumbre, en una prueba de hipótesis estadística hay DOS tipos de errores que uno puede cometer como resultado de una investigación:

RESULTADO DE UNA INVESTIGACION

	ACEPTAR	RECHAZAR
Ho: VERDADERA	OK	Error Tipo I
Ho: FALSA	Error Tipo II	OK

Supongamos sólo por un momento que en la realidad nuestra apreciación de la inteligencia de los estudiantes de biología esta sesgada por nuestra experiencia personal. Es decir, en realidad no existen diferencias reales con el resto de los estudiantes. Entonces, si tomamos una muestra de estudiantes al azar de la población de estudiantes de biología y estos estudiantes tienen un promedio de CI tal que es muy parecido al del resto de los estudiantes de la población, llevándonos a aceptar la Ho, entonces no hemos cometido ningún error. Si por el contrario y a pesar de que los estudiantes de biología son igualmente inteligentes que el resto, nuestra muestra correspondió a estudiantes sobresalientes y nos llevó a rechazar la hipótesis nula, entonces hemos cometido un error. Hemos concluido que los estudiantes de biología son más inteligentes cuando en la realidad no lo son. A este error lo denominamos Error Tipo I.

El **Error Tipo I** es la probabilidad que la muestra con media m realmente pertenezca a la población de media μ , pero que la hemos considerado como de una población diferente. Es decir, es la probabilidad de rechazar una hipótesis nula que es verdadera. La **región de aceptación** de la hipótesis nula corresponde a la región del Intervalo de Confianza.

Si por el contrario en nuestro ejemplo, los estudiantes de biología son efectivamente más inteligentes que el resto de los estudiantes universitarios (lo que todos sabemos) y nuestra *muestra* de estudiantes de biología tiene una media, \bar{Y} , tal que nos permite rechazar la H_0 , entonces no hemos cometido un error en la evaluación de la hipótesis. Sin embargo, si por simple azar hemos tomado una muestra de estudiantes de biología que tuvo en promedio un CI más bajo (\bar{Y} más bajo) que el real de la población de estudiantes de biología, lo que nos llevó a concluir, erróneamente, que ellos son igual de inteligentes que el resto, entonces hemos cometido un error. A este error lo llamamos Error Tipo II, que es la probabilidad de aceptar una hipótesis nula que es en verdad falsa.

Es muy importante tener presente que en casi todos los estudios científicos que leemos o que nosotros mismos desarrollamos, nunca sabemos si efectivamente cometimos un error o no. El cálculo de error no se trata de determinar si se cometió el error o no, sino de estimar la probabilidad de haber cometido tales errores. Puesto que en ciencias así como en otras actividades humanas mundanas asumimos una filosofía de escepticismo, no creemos a menos que nos presenten pruebas, le atribuimos mayor importancia al error Tipo I. Por ello y por razones prácticas, en prueba de hipótesis estadísticas del área de las ciencias básicas, fijamos la tasa de error Tipo I que consideramos aceptable y esta tasa de error aceptable la denominamos nivel de significancia.

El *Error Tipo I* tiene asociada una probabilidad **α** , la cual corresponde a la **Región de Rechazo** de la hipótesis nula. El valor de **α** es el **nivel de significancia** de una prueba estadística y se *fija* antes de realizar el experimento o estudio. En biología el valor al cual se fija α es normalmente 0.05 (5% o 1 en 20 veces). Esto es simplemente una convención y no hay ninguna razón matemática o nada mágico en el valor.

Puesto que las probabilidades deben sumar a la unidad y hemos dividido el espacio probabilístico en dos eventos (rechazo o aceptación), la **Región de Aceptación** contiene siempre una probabilidad acumulada de $1 - \alpha$.

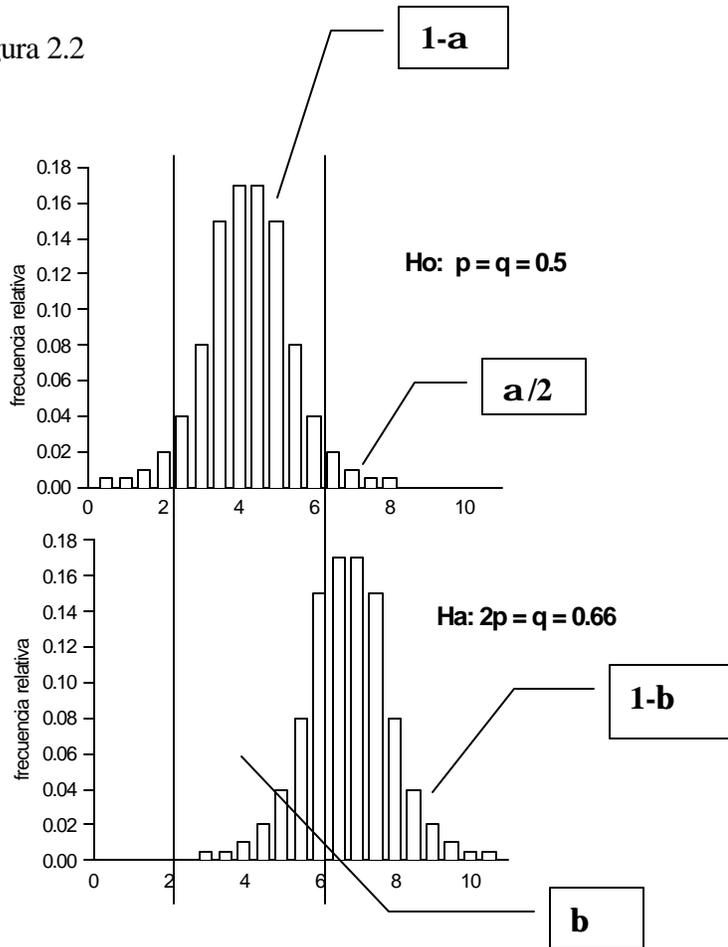
El *Error Tipo II* tiene asociada una probabilidad **β** . La región cuya probabilidad acumulada es igual a **β** corresponde a la zona de sobreposición entre la Región de Aceptación de la distribución especificada por la hipótesis nula y la distribución especificada por la hipótesis alternativa (cuya probabilidad es igual o menor a $1 - \alpha$).

Nuestra tasa de error Tipo I 'aceptable', significa que, por ejemplo, si ustedes realizan 40 pruebas estadísticas en el curso de su carrera científica, por simple azar se espera que se hayan equivocado al menos dos veces.

La relación entre la probabilidad de error Tipo I (α) y de error tipo II (β) se muestra en la Figura 2.2 para el caso de las distribuciones esperadas en número de hijas mujeres en familias con 8 hijos. El eje X está expresado en número de hijas mujeres en vez de proporción y al panel inferior deben por lo tanto restársele aproximadamente dos unidades (hijas). Para entender la

figura deben recordar que lo único que conocemos, es la probabilidad *esperada* bajo la hipótesis nula, es decir el panel superior.

Figura 2.2



El área por fuera de **b**, la región con una probabilidad acumulada de $1 - \mathbf{b}$, se llama el *poder* de la prueba estadística por cuanto mide la probabilidad de rechazar la hipótesis nula cuando esta es en efecto falsa.

¿Por qué no reducimos al máximo la probabilidad de cometer un error Tipo I (probabilidad de rechazar la hipótesis nula siendo verdadera), haciendo \mathbf{a} lo más chico posible?

- Podemos insistir en estar super seguros que no cometeremos un error y no rechazaremos esta hipótesis a menos que la probabilidad sea de 1% o menor. Por ejemplo, en criminología toda persona es inocente hasta que se pruebe lo contrario más allá de toda duda.

- Al hacer esto aumentamos la región de sobreposición con la distribución especificada por la hipótesis alternativa, aumentando la probabilidad **b** de cometer un Error Tipo II.
- Es decir, al disminuir el error Tipo I de rechazar la HIPOTESIS nula siendo verdadera aumentamos el error Tipo II de Aceptar la hipótesis nula debiendo rechazarla.
- Para un diseño experimental determinado, existe una relación inversa entre las probabilidades de cometer estos dos tipos de error en una prueba de hipótesis.

¿Cómo de terminamos cual Tipo de Error es más importante mantener bajo?

Como decíamos, en general, la filosofía de prueba de hipótesis es similar a la criminología: Es mucho más grave condenar a una persona inocente que dejar libre a un culpable.

En general, las consecuencias de aceptar una hipótesis nula que resulta ser falsa son menores que las consecuencias de rechazar una hipótesis nula que efectivamente es verdadera, puesto que al hacer esto estamos implícitamente aceptando la hipótesis alternativa. Esta es la filosofía de la “incredulidad” que domina el pensamiento científico.

Además, existe una razón más bien operacional para fijar la probabilidad de error Tipo I: es posible conocer y fijar *a priori* el nivel de probabilidad **a**. Este nivel de error depende exclusivamente de la distribución especificada por la hipótesis nula de no diferencias y es “análogo” (aunque con objetivos muy diferentes) al cálculo de intervalos de confianza. La probabilidad **b** de error Tipo II, al contrario, depende del valor de los parámetros especificados por la hipótesis alternativa. Normalmente NO conocemos estos valores y por ello es difícil *fijar* la probabilidad *a priori*.

En algunas circunstancias, tales como en aplicaciones industriales, conocemos exactamente el valor de los parámetros especificados por la hipótesis alternativa y entonces es posible también fijar un valor de **b**.

- En otras circunstancias, las consecuencias de aceptar la hipótesis nula cuando es falsa son claramente más graves que las consecuencias de rechazarla cuando es verdadera. Por ejemplo, cuando tratamos de evaluar el impacto ambiental de una industria que bota desechos tóxicos al ambiente, las consecuencias de concluir que los desechos NO tienen efectos cuando en verdad si los tienen pueden ser mucho más serias que el concluir erróneamente que los desechos si tiene efectos. En estas circunstancias, algunos autores sugieren modificar el error tipo I haciéndolo más grande (más permisible) y así disminuir el error tipo II.

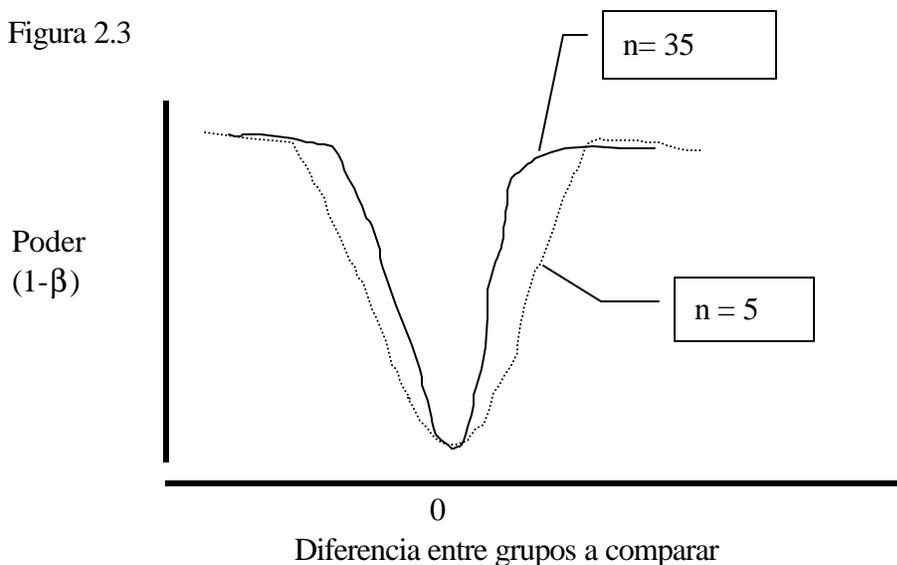
6. PODER DE UNA PRUEBA ESTADISTICA

Resulta obvio de la figura 2.2 que el poder de una prueba estadística depende del valor de parámetros especificados por la hipótesis alternativa. El área de superposición comprendida por $1-\beta$ varía dependiendo de la “localización” específica prescrita por la hipótesis alternativa. Para un diseño de muestreo o experimental determinado, es posible aumentar el poder de una prueba y mantener constante el Error Tipo I solamente a través de:

- Disminuir la varianza de la muestra y con ello la desviación estándar y de allí el error estándar
- Aumentar el tamaño de la muestra.

La relación entre el poder de una prueba estadística y el valor especificado por la hipótesis alternativa y el tamaño muestral se muestra en la figura 2.3:

Figura 2.3



- El cálculo del Poder de una prueba estadística no es un problema fácil. Resulta relativamente simple para los métodos estadísticos que nosotros veremos en esta clase, pero es mucho más complicado para métodos más sofisticados. Aún para estos métodos, hay bastante discusión acerca del tipo de errores que deben usarse en los cálculos (tasas *a priori* o *a posteriori*).
- En tiempos recientes ha habido una enorme discusión respecto de la necesidad de calcular el poder de todas las pruebas estadísticas que se usan en un estudio, para así poder juzgar si la falta de diferencias significativas se debe a que el poder de la prueba estadística es muy bajo.

- Más adelante veremos como calcular el poder de pruebas estadísticas simples cuando los factores son fijos o aleatorios.
- Nosotros nos vamos a mantener alejados de esta discusión por el momento y vamos a tratar de responder otra pregunta relacionada al poder de las pruebas estadísticas que normalmente los investigadores deben preguntarse ANTES de realizar un estudio:

¿Cuán grande debe ser el tamaño de la muestra (replicación) para un experimento?

Como mencionamos anteriormente, replicación aumenta la precisión de nuestras estimaciones y el poder de las pruebas de hipótesis, por lo que siempre resulta favorable incrementar el número de réplicas. Sin embargo los costos involucrados con la mantención de réplicas (o problemas éticos o legales) limitan el número de replicas que efectivamente se puede usar en un estudio.

Lamentablemente, frecuentemente no hay una respuesta simple a la pregunta de cual es el nivel de replicación adecuado para un estudio y sin una estimación de la variabilidad de la variable respuesta No hay ninguna respuesta que pueda darse. El método elemental para determinar el número de replicación adecuado se basa en la prueba de hipótesis acerca de diferencias entre medias de grupos.

Para dos muestras y usando la prueba normal estándar, se puede desarrollar una prueba simple para la prueba de diferencias entre las dos medias: $\delta = \mu_i - \mu_j$ con una varianza común y conocida σ^2 . El método determina el número de replicas necesario para someter a prueba la hipótesis de diferencias δ entre las medias con un determinado error Tipo I y Tipo II.

El número de replicas necesario esta determinado por:

- La varianza σ^2
- La magnitud de las diferencias δ
- El nivel de significancia α , o probabilidad de error Tipo I
- El poder de la prueba 1-B, o probabilidad de error Tipo II.

El nivel de replicación necesario para cada grupo, r, para pruebas de dos colas es:

$$r \geq 2[z_{\alpha/2} + z_{\beta}]^2 \left(\frac{s}{d} \right)^2$$

En donde z corresponde al valor de la distribución normal estandarizada evaluada en $\alpha/2$ y β , respectivamente. Es posible estimar el nivel de replicación al conocer el coeficiente de variación %CV y sustituirlo por σ en la ecuación y expresando la diferencia δ como un porcentaje de la

gran media de los grupos. Esto evita tener que conocer exactamente el valor de las medias esperadas, pero es necesario determinar el porcentaje de diferencias esperadas o que se desea determinar. Aún así, determinar el parámetro σ/δ , denominado “parámetro no central” no es fácil.

Los valores calculados son estimaciones y aproximaciones. Estos son normalmente determinados sobre la base de estimaciones de varianzas de experimentos piloto o experimentos similares anteriores. Si se usan estimaciones basadas en experimentos anteriores, se debe estar dispuesto a asumir que las condiciones de esos estudios eran similares a las condiciones del estudio que se va a realizar, de manera que las varianzas esperadas en la variable respuesta sean efectivamente similares.

¿Qué características son deseables en una prueba estadística?

No existe realmente una prueba estadística *ideal* pues siempre estamos jugando con estos compromisos de cometer errores al juzgar los resultados de un experimento.

Una prueba estadística ideal debería:

1. Ser CONSERVATIVO: Tener una probabilidad baja de error Tipo I.
2. Ser PODEROSO: Tener una probabilidad baja de Error Tipo II.
3. Ser ROBUSTO: Tener baja sensibilidad a desviaciones de los supuestos de la prueba.
(en general, mientras menor sea el número de supuestos de una prueba, menor es el poder)

Clase 3: ANDEVA, Principios Básicos, Tabla de ANDEVA y Supuestos

1. Métodos Estadísticos de Prueba de Hipótesis

Existe un sin número de métodos estadísticos de prueba de hipótesis. Muchos de ellos están diseñados para diferentes circunstancias, diferentes tipos de datos, diferentes hipótesis específicas, etc.

- Una de las clasificaciones o divisiones tradicionales en los métodos de prueba de hipótesis ha sido entre los llamados métodos paramétricos y los mal llamados métodos ‘no paramétricos’.
- **Métodos paramétricos** son aquellos métodos estadísticos que realizan inferencias acerca de parámetros de una población, tales como media y varianza, y que *suponen* una determinada distribución de los valores de la variable bajo estudio (ej. normalidad).
- Los métodos mal llamados **no paramétricos** o más bien de **‘libre distribución’** no dependen estrictamente de supuestos acerca de la *distribución* de la variable bajo estudio. Esto NO significa que sean libres de supuestos. Todos estos métodos, en mayor o menor grado, tienen supuestos acerca de las propiedades de la variable bajo estudio.
- En general los métodos no paramétricos realizan inferencias acerca de la mediana de una población y no de la media y son los únicos métodos apropiados cuando los datos son categóricos, no ordenables y casi los únicos apropiados cuando los datos son expresados como rankings.
- Sin embargo, los métodos no paramétricos no permiten diseños experimentales complejos y en general tienen menor poder que métodos paramétricos equivalentes, cuando estos últimos pueden aplicarse.

Todo esto para decirles que nosotros no veremos métodos ‘no paramétricos’ en detalle en esta clase y nos concentraremos en el método paramétrico de mayor uso en ciencias experimentales: el Análisis de Varianza. Veremos un par de pruebas de distribución libre para comparar medianas, las que son útiles para analizar resultados de diseños experimentales simples.

Como vimos en el curso introductorio BIO 242A, el análisis de varianza es por mucho el método estadístico más usado en ciencias biológicas ambientales y si consideramos la prueba t-Student como un caso especial de ANDEVA, entonces esta es la técnica de análisis estadístico más usada en ciencias en general.

La popularidad de ANDEVA tiene que ver con la facilidad de tratamiento matemático, y lo intuitivo que es su aplicación a situaciones reales. Además, ANDEVA es uno de los métodos de prueba de hipótesis más poderosos (gran capacidad de rechazar una hipótesis nula que es falsa).

Otra ventaja es que ANDEVA no es solamente útil como prueba de hipótesis, sino que al contrario de otras técnicas estadísticas, también entrega información acerca de los factores de variación en la naturaleza. ANDEVA permite analizar diseños de muestreo y experimentales extremadamente complejos y entender como y que factores contribuyen a mantener o a aumentar la variabilidad en una variable determinada.

La técnica fue desarrollada por Ronald A. Fisher, un estadístico inglés de principios de siglo que desarrolló la técnica para aplicarla a estudios en agricultura. Muchas de las pruebas y diseños originales fueron pensados en experimentos manipulativos típicos de estudios de agricultura y algunos nombres aún permanecen en la jerga estadística. Hoy en día el método se usa en casi todas las ramas de la ciencia y tecnología, incluyendo psicología y medicina.

El método examina diferencias entre medias de distintos grupos, a través de comparar varianzas dentro, versus entre grupos.

- Por ejemplo:
 - Comparar la altura de hombres y mujeres de este curso.
 - Comparar el efecto de una droga en cuatro diferentes razas de chanchitos
 - Comparar la densidad de ratones en laderas sur versus laderas norte
- La *comparación entre medias* de la o las poblaciones se realiza a través de *analizar las varianzas* de diferentes muestras de estas poblaciones o grupos.

Como toda prueba de hipótesis estadística, ANDEVA tiene una serie de supuestos acerca de la distribución de los datos o las propiedades de las varianzas. Aquí veremos primero el análisis mismo y luego volveremos sobre los supuestos.

2. HIPOTESIS EN ANALISIS DE VARIANZA

En TODOS los ANDEVA se somete a prueba Hipótesis de la forma:

Ho: $\mu_1 = \mu_2$ (hipótesis de dos grupos. Puede usarse una prueba de t)

Ha: $\mu_1 \neq \mu_2$

a = 2 (**a** = grupos que se desea comparar)

Ho: $\mu_1 = \mu_2 = \mu_3 = \mu_4$

Ha: Al menos un μ_i es diferente
a = 4

El 'estándar de comparación' usado por la prueba de ANDEVA que nos permite evaluar la probabilidad de que los datos observados se produzcan por simple azar, es la denominada Distribución de F (en honor a Fisher).

3. LA DISTRIBUCION DE F

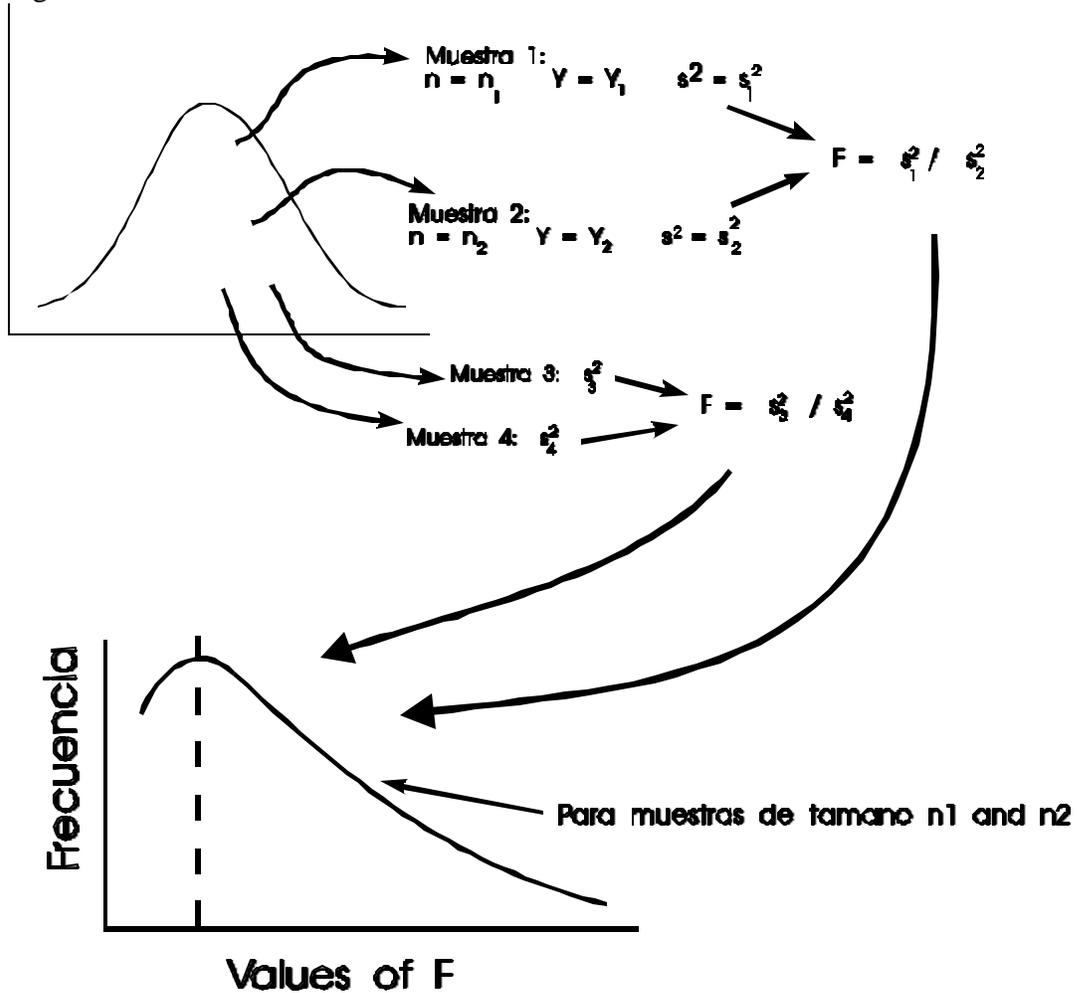
Imaginen el siguiente ejercicio:

De una Población conocida tomamos dos muestras aleatorias consecutivas de *tamaño* **n1** y **n2** y a cada muestra le calculamos la media ($\bar{Y} = m$) y la varianza (s^2)

si $y \sim N(\mu, \sigma^2)$ y m_1 y m_2 son dos muestras aleatorias de la población, entonces la razón de las varianzas s_1^2 y s_2^2 tiene una distribución F con $v_1 = n_1 - 1$ y $v_2 = n_2 - 1$ grados de libertad.

- Existe una curva de F para cada combinación de grados de libertad del numerador y denominador. Esto significa que la forma de la curva varía con el tamaño muestral y debemos determinar el valor de probabilidad desde la curva apropiada.
- Los grados de libertad del numerador y denominador son aquellos con que se calculan las varianzas respectivas (usualmente $n - 1$, pero en otros casos se usa una expresión diferente) de las muestras tomadas de la población
- Todo valor de F tiene dos fuentes de grados de libertad (v_1 y v_2), y siempre se deben reportar ambas fuentes ($F_{[v_1, v_2]}$) o no es posible obtener la probabilidad asociada a ese valor de F. 'Se ha visto' en algunas publicaciones valores de F a los cuales se asocia un solo valor de grados de libertad, lo cual es claramente un error.

Figura 3.1



La razón de las varianzas de muestras de una población con distribución normal siguen una distribución de F. Puesto que estas varianzas provienen de la misma población normal, la razón de las varianzas deberá ser cercana a uno.

- Cada curva de F nos muestra la *probabilidad* de encontrar diferencias entre dos varianzas por simple azar.
- La distribución de F también puede producirse por dos poblaciones con diferentes medias pero las mismas varianzas.

- En esta distribución de F tenemos entonces un **estándar** contra que comparar resultados observados. Esta distribución nos permitiría entonces determinar un nivel de significancia para decir cuando dos varianzas son significativamente diferentes.

4. ANALISIS DE VARIANZA

1. Imaginen que estamos interesados en saber el efecto subletal de depredadores sobre insectos acuáticos. Nuestro modelo biológico sugiere que especies acuáticas de presa que han coexistido con depredadores específicos sobre tiempo evolutivo, pueden poseer receptores específicos que les permiten detectar la presencia de depredadores. Específicamente, queremos demostrar si la concentración de la proteína involucrada en los receptores de las sustancias (exudados) emitidas por depredadores acuáticos son más altas cuando existen señales químicas en el agua (presencia de depredadores) que cuando los depredadores están completamente ausentes.

Entonces, diseñamos un experimento lo más simple posible en cual ponemos 20 insectos en acuarios individuales y luego asignamos 10 de estos acuarios a una condición 'control', sin depredadores y los otros diez al tratamiento 'con exudado de depredadores'. Esta estructura constituye nuestro *diseño de tratamientos*, el cual puede ser inadecuado para controlar por todas las fuentes potenciales de error que podemos introducir en el experimento. La selección correcta del diseño de tratamientos depende de los tratamientos seleccionados, pero para efectos de este ejemplo supondremos que no existen otras fuentes de error importantes.

Nuestro tratamiento consiste de dos niveles:

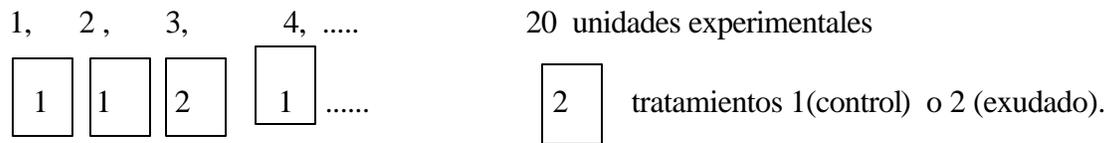
A=1: Con exudado (químicos) de depredadores en el agua

B=2: Sin exudados de depredadores en el agua

Es decir, tenemos **dos grupos (a = 2)** de 10 individuos cada uno, a los cuales aplicaremos los tratamientos correspondientes y luego de un periodo de exposición determinado los sacrificaremos y mediremos la concentración de la proteína involucrada en los receptores.

Ahora debemos decidir que diseño experimental nos permite controlar adecuadamente el error experimental (y obtener una estimación de la magnitud de dicho error) y a la vez incorporar el diseño de tratamientos. El diseño más simple posible que permite controlar por error experimental, garantizando independencia de las unidades experimentales, es la asignación de tratamientos completamente aleatoria a las unidades experimentales. Esto significa que los niveles del tratamiento (en este caso sin exudado o con exudado) son aplicados en forma completamente aleatoria a cada una de las unidades experimentales:

El diseño se ve así:



Este diseño experimental, en cual los tratamientos son aplicados a cada unidad experimental en forma completamente aleatoria se conoce como: **Diseño Completamente Aleatorio o “CRD”**.

Un error común en este tipo de experimentos es poner todos los individuos bajo un determinado tratamiento en un solo acuario. Esto obviamente reduce el número de acuarios a dos y con ello los costos de montaje y mantención. Sin embargo, aún y cuando los insectos destinados a cada acuario hayan sido seleccionados al azar, el ponerlos juntos viola uno de los supuestos fundamentales de todo análisis estadístico, esto es la “independencia de las unidades experimentales”. Esto significa que en lugar de tener 10 réplicas de cada nivel de 1 tratamiento, en la realidad NO tenemos réplicas. Daría lo mismo si tuviéramos ahora solamente un insecto por tratamiento, pues si cualquier factor ajeno a los tratamientos afecta a un acuario más que al otro (luz, temperatura del agua, oxígeno, etc), afectará a todos los insectos de ese acuario. Entonces, la idea de asignar los tratamientos al azar es garantizar la falta de sesgo, peor también la independencia de estas unidades.

2. Nuestras hipótesis son:

a. Expresadas en forma biológica:

Ho: No existen diferencias en la concentración de proteína receptora en aquellos insectos en la presencia versus aquellos en ausencia de señales o exudados de depredadores

Ha: Existen diferencias significativas en la concentración de proteínas entre estos dos grupos experimentales, la cuales son atribuibles a la presencia y ausencia de exudados de depredadores

b. En forma estadística:

$$Ho: \mu_1 = \mu_2$$

$$Ha: \mu_1 \neq \mu_2$$

3. Los grupos que queremos comparar tienen tamaños de muestra iguales, $n_1 = n_2$

4. Ahora podemos razonar de la misma manera que lo hicimos al tratar de comprobar si una moneda esta sesgada o si los estudiantes de biología tienen un coeficiente intelectual más alto que

el resto. Lo que hacemos es especificar cuales serían los resultados *esperados* si la H_0 fuera verdadera, es decir que las diferencias observadas entre los dos grupos se deban al simple azar.

5. Para “generar” los resultados esperados y siguiendo a Fisher, nos valemos de la distribución esperada de la razón de varianzas de muestras de una misma población normal. Para hacerlo entonces debemos calcular dos varianzas de muestras de una misma población y luego calcular sus razones. Esto requiere calcular las sumas de cuadrados de las desviaciones entre las observaciones y sus medias de la siguiente manera:

A. Primero que nada calculamos las **suma de los cuadrados de las desviaciones (SC)** de cada uno de los grupos que queremos comparar (A y B) usando la media muestral respectiva de cada grupo.

Las SC de las desviaciones dentro del grupo $k = 1$ serán:

$$SC_{dentro} = \sum_{i=1}^{n_1} (y_{ik} - \bar{Y}_k)^2 \quad SC_{dentro} = \sum_{i=1}^{n_2} (y_{ik} - \bar{Y}_k)^2$$

k=1 k=2

Luego computamos un promedio de estas SC de las desviaciones dentro de cada grupo.

Si la H_0 es verdadera, las varianzas dentro de cada grupo son una estimación de la misma varianza poblacional y el promedio de estas varianzas es también una estimación de la varianza total poblacional. Este promedio de desviaciones cuadradas dentro de los grupos se llama:

SUMA DE CUADRADOS DENTRO DE GRUPOS (= WITHIN)

$$SC_{dentro} = \sum_{k=1}^a \sum_{i=1}^n (Y_{ik} - \bar{Y}_k)^2$$

Si $n_1 = n_2 = n$,

entonces simplemente dividimos este termino por $a(n-1)$, los grados de libertad para calcular la varianza dentro de grupos.

C. También podemos calcular una varianza entre las medias de los grupos. Para esto calculamos la **suma de cuadrados de las desviaciones de las medias entre grupos**.

SUMA DE CUADRADOS ENTRE GRUPOS (= BETWEEN)

$$SC_{entre} = \sum_{j=1}^a (\bar{Y}_j - \bar{\bar{Y}})^2$$

para transformar esta suma de cuadrados en la varianza, dividimos este termino por **a-1**, donde a es el numero de grupos de medias sobre el cual el calculo esta basado.

Si la Ho es verdadera, el promedio de estas varianzas entre las medias será otra estimación de la misma varianza poblacional.

B. Finalmente, podemos fácilmente calcular la **varianza total** de todas las observaciones, o la Suma de Cuadrados Total de las desviaciones de las observaciones con respecto a la gran media:

SUMA DE CUADRADOS TOTAL

$$SC_{total} = \sum_{ik}^n (y_{ik} - \bar{\bar{Y}})^2$$

D. Puesto que las Sumas de Cuadrados son *Aditivas*, podemos calcular la suma de cuadrados total a partir de la expresión:

$$SC_{Total} = SC_{dentro} + SC_{entre}$$

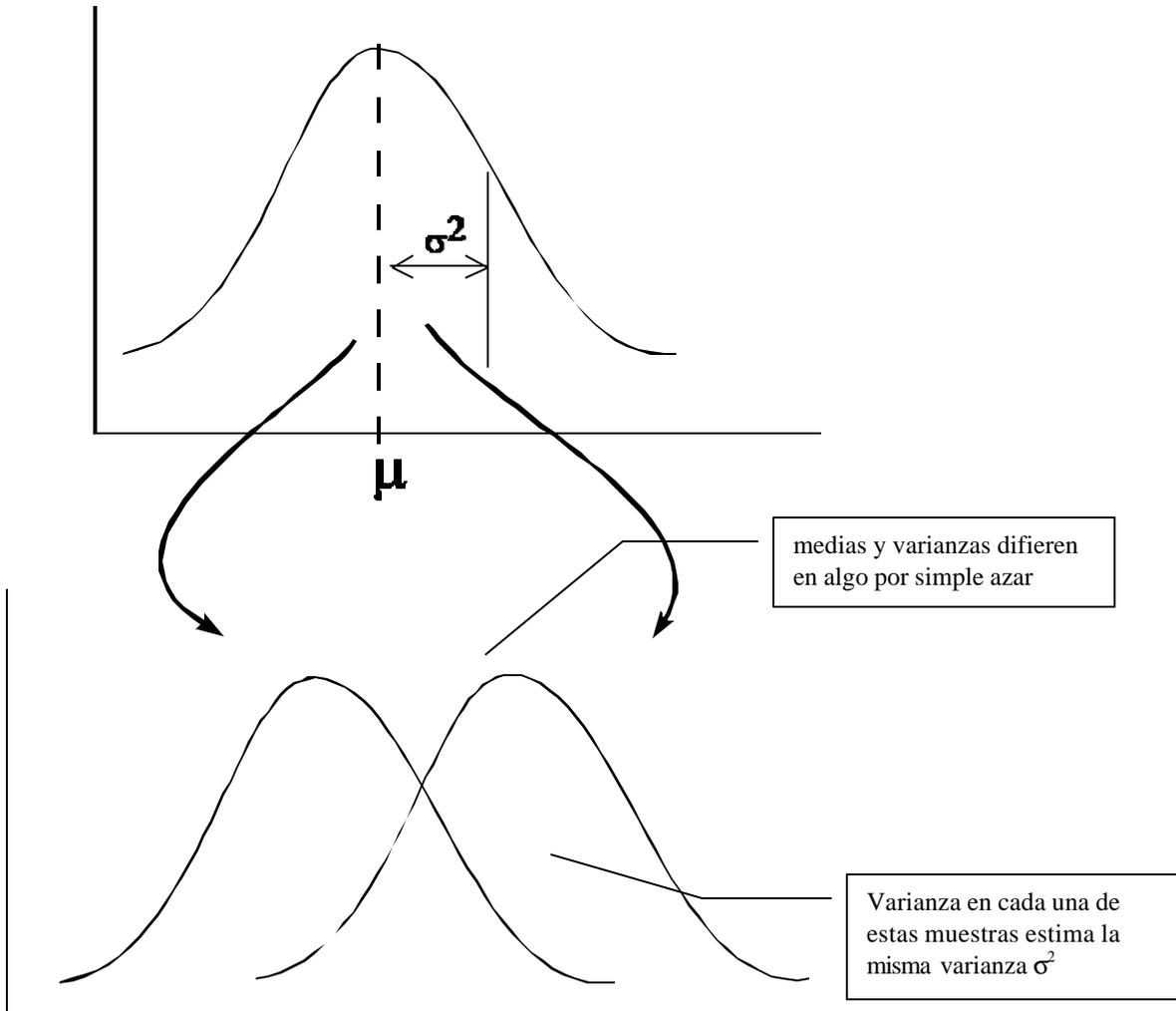
$$SC_{total} = \sum_{ik}^n (y_{ik} - \bar{\bar{Y}})^2 = \sum_{k=1}^a \sum_{i=1}^n (Y_{ik} - \bar{Y}_k)^2 + \sum_{j=1}^a (\bar{Y}_j - \bar{\bar{Y}})^2$$

En la practica, usando el principio de aditividad de las sumas de cuadrados, uno calcula la suma de cuadrados total y la suma de cuadrados dentro de grupos, y por diferencia se calcula la suma de cuadrados entre grupos.

La figura 3.2 muestra las distribuciones esperadas de muestras de una misma población.

Figura 3.2

H₀: Muestras de la MISMA población



E. ¿Cómo podemos comparar estas dos varianzas y saber si las diferencias observadas son producto de simple azar?

Si las dos varianzas: la varianza dentro de grupos y la varianza entre grupos efectivamente provienen de la misma población, entonces podemos comparar la razón entre las varianzas contra lo que se espera en una distribución de F.

$$F_{[v1,v2]} = \frac{\text{Varianza Entre Grupos}}{\text{Varianza Dentro de Grupos}}$$

- Si efectivamente s^2 entre grupos y s^2 dentro de grupos son estimadores de la misma σ^2 , entonces el valor de F observado deberá ser cercano a 1.
- Si los dos grupos son diferentes entonces la s^2 entre grupos será más grande que la s^2 dentro de los grupos y F será mucho mayor que 1.

F. ¿Cómo podemos determinar la significancia de nuestro valor de F?

Comparamos el valor de F observado con el valor F esperado por simple azar en una tabla de F, y seguimos nuestro criterio definido *a priori* de nivel de significancia al 5%.

Si la probabilidad de obtener un F mayor o igual al observado es menor que 0.05, decimos que las medias de los dos grupos **NO** provienen de la misma población. En otras palabras, los dos grupos NO tienen la misma media.

6. LA TABLA DE ANDEVA

Ya hemos calculado la SUMA de CUADRADOS de los términos que necesitamos para una ANDEVA

Sabemos que si los grupos (a_1 y a_2 en nuestro ejemplo) provienen de la misma población y por ende tienen la misma media (μ), es decir SI la H_0 es verdadera, entonces la varianza promedio DENTRO de los grupos será similar a la varianza promedio ENTRE los grupos

Ahora necesitamos dividir las SUMAS de CUADRADOS de las desviaciones por los grados de libertad correspondientes para así obtener una estimación de estas varianzas. Estas estimaciones de varianzas en la ANDEVA se llaman CUADRADOS MEDIOS (CM). Como veremos más adelante, a excepción de la varianza del error experimental o residual, el resto de los cuadrados medios no son exactamente varianzas, sino que la sumatoria de varios componentes de varianza.

¿Cómo sabemos cuantos grados de libertad utilizar?

Los grados de libertad son igual al número de observaciones sobre las cuales se basó el cálculo de la suma de cuadrados menos uno. En realidad, como una regla general, es el número de observaciones independientes menos el número de parámetros que necesitamos estimar para ese cálculo, es decir el número de observaciones libres de variar. En el caso de la suma de

cuadrados entre grupos, nuestra estimación de varianza esta basada en a grupos (en el ejemplo de efectos subletales $a = 2$) y además debemos conocer la gran media de todos los grupos (Y-barra-barra), de manera que debemos restar un grado de libertad a la estimación de varianza entre grupos ($a-1$). Recuerden que no es necesario conocer la media de cada uno de los grupos separados para obtener la gran media puesto que basta calcular el promedio de todas las observaciones y dividir por $\sum n_i$

TABLA DE ANDEVA

Fuente de Variación	g.l.	SC	CM	F obs	P
ENTRE GRUPOS	$a-1$	SCe	$CM_e = SCe/(a-1)$	$F_{obs} = CM_e/CM_d$	
DENTRO DE GRUPOS (Error Experimental)	$a(n-1)$	SCd	$CM_d = SCd/a(n-1)$		
TOTAL	$an-1$	SCtotal			

g.l.= Grados de Libertad

En esta tabla **P** = probabilidad de encontrar un valor de F así tan grande por simple azar. Mientras más grande es el valor de F, mayor es la contribución de la varianza entre grupos (efecto del tratamiento) en relación a la varianza dentro de grupos o varianza experimental. Si la probabilidad de encontrar un F así tan grande es < 0.05 (nuestro nivel de significancia pre-establecido) entonces decimos que: *“la concentración de proteínas receptoras es significativamente mayor (o menor) en individuos expuestos a exudados de depredadores”*

Ahora bien, si los tamaños muestrales NO son iguales entre todos los grupos, entonces el valor de “n” debe ser estimado usando la siguiente ecuación:

$$n_0 = \frac{1}{a-1} \left(\sum_{j=1}^a n_i - \frac{\sum_{j=1}^a n_i^2}{\sum_{j=2}^a n_i} \right)$$

El valor de n_0 es cercano, pero siempre un poco menor que el promedio de los tamaños muestrales, a menos que todas las muestras sean del mismo tamaño. Cuando el número de

observaciones independientes (tamaño muestral) o “**REPLICAS**” es diferente en los distintos grupos a comparar, el diseño de ANDEVA se llama **desbalanceado**.

La gran mayoría de los programas estadísticos disponibles para PC o Macintosh manejan diseños desbalanceados sin problemas, al menos para prueba de hipótesis. Los cálculos de sumas de cuadrados son un poco diferentes a las ecuaciones que vimos aquí.

7. SUPUESTOS DE ANDEVA

Antes de seguir adelante con distintos diseños experimentales, necesitamos volver a los supuestos de análisis de varianza.

TODAS LAS PRUEBAS DE HIPOTESIS ESTADISTICAS SUPONEN:

1. Muestreo Aleatorio

- Al realizar el muestreo de una población, el *diseño de muestreo* debe ser *aleatorio*. Un muestreo no aleatorio introducirá un sesgo en nuestros datos y este sesgo **NO** tiene solución.
- Así mismo, la asignación de réplicas a distintos tratamientos de un experimento debe ser al azar

2. Independencia de los Errores.

- Esto significa que aquellos factores que tienen un efecto sobre la variable a analizar (ej. efecto del tamaño sobre tasas metabólicas) pero que **NO** forman parte del o los tratamientos bajo estudio, **NO** deben estar *correlacionados* con nuestros tratamientos.
- Falta de independencia de errores también se produce cuando las replicas (acuarios, individuos, etc.) se afectan unas a otras.
- La única manera de garantizar independencia de errores es tomar muestras o asignar réplicas completamente al azar. La falta de independencia de errores en un diseño **NO** tiene solución.
- Cuando el número de réplicas de un experimento es bajo, se debe tener cuidado que no queden todas las réplicas de un mismo tratamiento no queden agrupadas. Esta es la idea de interspersión de la réplicas (Hurlbert).

ANALISIS DE VARIANZA ADEMÁS TIENE ESTOS SUPUESTOS:

3. Normalidad.

- ANDEVA requiere que nuestros datos estén aproximadamente normalmente distribuidos

$\sim N(\mathbf{m}, \mathbf{s}^2)$

- Existen muchas maneras de verificar si nuestros datos cumplen o no con este supuesto. En el curso Bio-242A vimos algunas maneras gráficas de someter a prueba este supuesto y mencionamos algunas pruebas de bondad de ajuste que comparan la distribución de los datos observada contra la distribución esperada (generada) basados en la media y varianza de la muestra.
- Cuando nuestros datos NO cumplen con este supuesto, usualmente es posible *transformar* los datos usando una transformación no lineal (ej. logaritmos, seno, arcoseno, etc.). Luego de la transformación los análisis deben realizarse sobre los datos transformados.
- Si nuestros datos están normalmente distribuidos, entonces los ‘errores’ (variación no explicada) deben estar también normalmente distribuidos. Es decir, si $y \sim N(\mu, \sigma^2)$, entonces $y - \mu = \text{residuos} \sim N(0, \sigma^2)$
- La mayoría de las investigaciones muestran que ANDEVA es robusto a desviaciones de normalidad, incluso desviaciones relativamente importantes cuando los tamaños muestrales (replicación) son grandes.
- Bajo determinadas circunstancias, en caso de desviaciones extremas de normalidad, es posible transformar los datos a rankings y realizar un ANDEVA sobre estos rankings. Esto es equivalente a pruebas no paramétricas o de distribución libre, pero tiene alguna restricciones. Más adelante veremos un poco más de esta alternativa.

4. Homogeneidad de Varianza (Homosedasticidad).

- Este es el supuesto más importante de ANDEVA porque es más sensitivo a estas desviaciones que a desviaciones de normalidad.
- Puesto que la *varianza dentro de grupos* se supone que estima la misma varianza poblacional que la *varianza entre grupos* (si H_0 es verdadera), Todos los grupos a comparar deben tener aproximadamente la misma varianza.
- Existen transformaciones que solucionan este problema de varianzas heterogéneas y al mismo tiempo pueden solucionar desviaciones de normalidad.

- ANEDVA es también relativamente robusto a desviaciones de homogeneidad de varianza y muchas de las pruebas estadísticas usadas para verificar este supuesto son demasiado sensibles. Como dice Tony Underwood, es como meterse al agua en un bote a remos para ver si el mar esta suficientemente tranquilo para que navegue un trasatlántico.
- La mayoría de las pruebas estadísticas no paramétricas o de distribución libre también son sensibles a desviaciones de homogeneidad de varianza. De hecho, la mayoría son tan sensibles como ANDEVA a estas desviaciones.

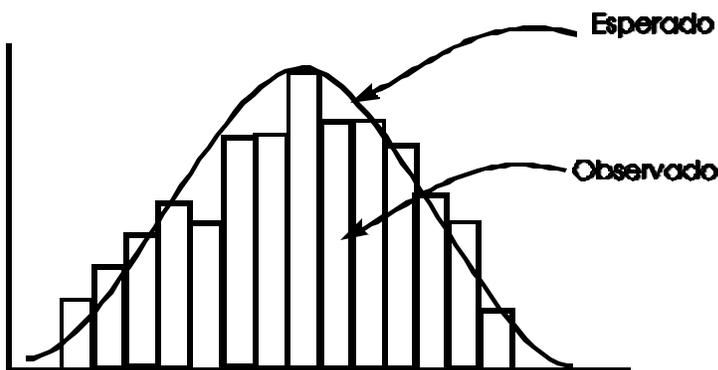
¿Qué significa que una prueba estadística sea sensible a normalidad u homogeneidad de varianza?

- Cuando una prueba estadística es sensible a desviaciones de los supuestos, significa que si realizamos una ANDEVA con datos que presentan varianzas heterogéneas, no podemos estar seguros que el Error Tipo I sea realmente 0.05.
- En general desviaciones de los supuestos llevan a tasas infladas de error tipo I. El problema es que no sabremos exactamente cuales son las probabilidades de error tipo I en el estudio.

8. COMO VERIFICAR NORMALIDAD DE LOS DATOS

Ajustando una curva normal.

Como mencionamos anteriormente, es posible verificar si nuestros datos se ajustan a la distribución esperada bajo el supuesto de normalidad a través de usar la ecuación que describe la curva normal (Z). Con esta ecuación y usando los la desviación estándar y media de la muestra como estimadores de σ y μ , se pueden generar los valores de frecuencia esperados y luego, la distribución de datos observados se compara a la distribución esperada (normal) usando una prueba de bondad de ajuste.

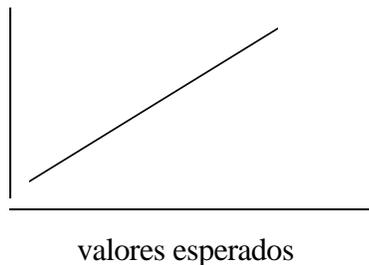


Existen Muchas pruebas de bondad de ajuste. Entre las más usadas para este tipo de comparaciones es la prueba de Kolmogorov-Smirnov.

2. Método gráfico.

Conociendo la desviación estándar y la media poblacional (o sus estimadores) podemos calcular las frecuencias acumuladas esperadas de una distribución normal. De esta manera obtendremos un valor esperado de probabilidad acumulada por cada observación. Luego, si graficamos nuestros datos contra estos valores esperados debemos observar una línea recta, *si* los datos siguen una distribución normal. Desviaciones de la línea recta indican desviaciones de normalidad.

Esto es lo que básicamente hace un gráfico Quantile-Quantile o Q-Q Plot o un gráfico de Z-scores. El gráfico Q-Q grafica los valores de la variable en forma ordena en el eje Y y los valores de los cuantiles de una distribución teórica determinada en el eje X. En este caso usamos la distribución normal estandarizada para generar las distribuciones esperadas. Si nuestros datos siguen una distribución normal, entonces los puntos deberían estar en una línea recta con intercepto igual a μ y pendiente igual a σ



3. Transformaciones de Datos.

- Las transformaciones lineares cambian la posición de nuestra distribución, pero No cambian la forma de la distribución.

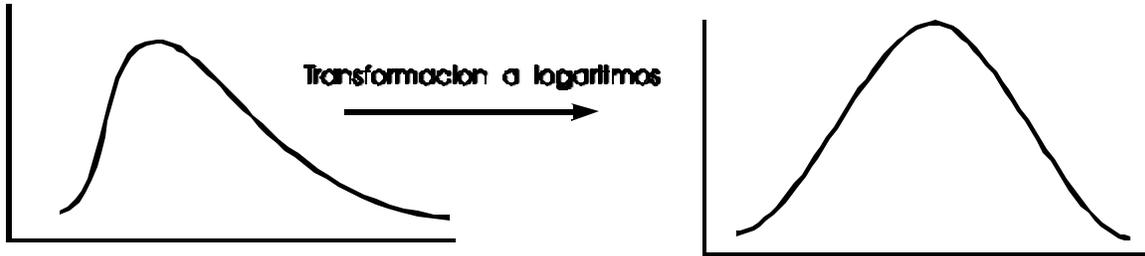
Si $y \sim N(\mu, \sigma^2)$, entonces $y - 8 \sim N(\mu - 8, \sigma^2)$

Si $y \sim N(\mu, \sigma^2)$, entonces $y - \mu \sim N(0, \sigma^2)$

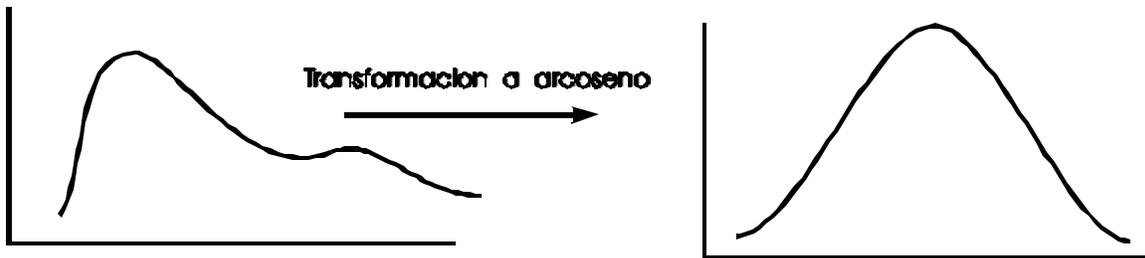
- Las transformaciones no lineares cambian la escala de referencia de los datos y así cambian la forma de la distribución.
- Los datos transformados son aún los mismos, solamente están expresados en una escala diferente.

Ejemplo de Transformaciones:

Logaritmos:



ArcoSeno (arcoseno de la raíz cuadrada de y):



Luego de las transformaciones es necesario nuevamente verificar que los datos transformados efectivamente cumplan con el supuesto de normalidad.

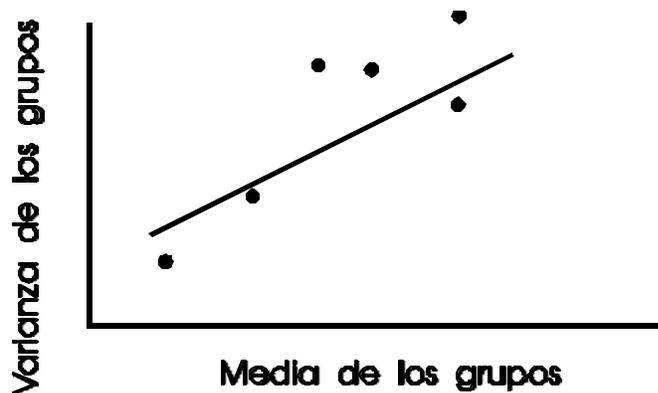
El supuesto es que CADA grupo tiene una distribución normal, NO todos los grupos de datos crudos juntos. El principal problema para verificar normalidad de los datos, especialmente en estudios experimentales es la cantidad de datos. Tanto estimaciones visuales como pruebas estadísticas son difíciles de aplicar cuando existen pocas observaciones. Lo más común es verificar normalidad en los residuos (errores que quedan luego de ajustar el modelo). Esto permite aumentar el tamaño muestral sobre el cual se puede evaluar normalidad. Por ejemplo, si queremos comparar tres grupos con medias Y_1 , Y_2 y Y_3 que estiman las medias poblacionales

μ_1 , μ_2 y μ_3 , entonces podemos restar la media a cada grupo a las observaciones de ese grupo, agrupar todas las observaciones (todos los grupos ahora tendrán media cero) y calcular la normalidad de las desviaciones (residuos) resultantes.

Afortunadamente, Análisis de Varianza es muy robusto al supuesto de normalidad. Esto significa que desviaciones de normalidad, aún con bajos números muestrales, no afectan grandemente los resultados e interpretación de los análisis. Las distribuciones que parecen causar los mayores problemas en ANDEVA son las distribuciones multimodales (varias modas en cada grupo a comparar). Muchas veces es posible cambiar nuestro modelo explicativo y rediseñar el muestreo de manera que cada moda represente ahora un grupo distintivo.

9. COMO VERIFICAR HOMOGENEIDAD DE VARIANZA DE LOS DATOS

1. *Método Visual.* Si hay varios grupos (> 3) se puede observar si existe heterogeneidad de varianza si existe una *correlación (positiva) entre la media y la varianza*. La existencia de correlación entre media y varianza es indicación de que las varainzas no son homogeneas a través de los grupos. Esta apreciación es posible solamente cuando se tienen varios grupos a comparar.



2. *Aplicar una prueba estadística para someter a prueba la hipótesis:*

$$H_0: s_1^2 = s_2^2 = s_3^2$$

Existen varias pruebas estadísticas. Las mas usadas son:

- a) Test de Bartlett
- b) Test C de Cochran
- c) Test de Fmax de Hartley.
- d) Test de Levene
- e) Test de Scheffé

Una de las pruebas más usadas en la literatura biológica es la prueba de Bartlett. Desafortunadamente, bajo muchas circunstancias esta prueba indica que hay desviaciones significativas de homogeneidad de varianzas en circunstancias que el grado de heterogeneidad no causa excesivo error Tipo I en un análisis de varianza. Por ello, y porque esta prueba además es sensible a desviaciones de normalidad, muchos autores han recomendado no usarla.

La prueba de Fmax que vimos en el curso Bio-242A es atractiva por cuanto es muy fácil de aplicar. Sin embargo, esta prueba también entrega tiende a indicar la existencia de varianzas heterogéneas cuando la variación más pequeña observada es chica, aún y cuando las otras varianzas sean iguales. Además, la prueba de Fmax también es sensible a desviaciones de normalidad.

La prueba C de Cochran es probablemente la más útil. La prueba también es sensible a desviaciones de normalidad en los datos, pero menos que otras pruebas propuestas. Aquí incluimos esta prueba y la prueba de Fmax

Test Fmax de Homogeneidad de Varianzas

- a. Encontrar la varianza más grande y la más chica de todos los grupos a comparar
- b. Calcular la razón de estas varianzas: esta es la Fmax
- c. Elegir un error (α) aceptable de cometer errores (Tipo I), normalmente 0.05
- d. El test supone que todos los grupos a comparar son del mismo tamaño. Si los grupos son distintos, usar el grupo con menor tamaño muestral (n) para calcular los grados de libertad (aunque este no sea el con menor o mayor varianza).
- e. Mirar en la tabla de distribución acumulada de Fmax los valores de probabilidad con **a** número de grupos y n-1 grados de libertad (tamaño muestral más chico).
- f. Si *Fmax observado* mayor que el de la tabla a un $\alpha = 0.05$, entonces rechazamos la hipótesis nula que las varianzas son homogéneas (= existe heterogeneidad e varianzas)

Test C de Cochran

- a. Encontrar la varianza más grande de todas los grupos a comparar (s^2_{mayor})
- b. Calcular la sumatoria de las varianzas de todos los grupos
- c. Calcular la razón, C de Cochran, de estos dos estadísticos:

$$C \text{ de Cochran} = \frac{s^2_{mayor}}{\sum_{i=1}^a s^2_i}$$

- d. Comparar el valor de C observado con valores tabulados en una tabla de C de Cochran con a grupos y n-1 grados de libertad en cada muestra

4. *Transformaciones.*

- Las transformaciones no-lineares también tienen efecto sobre el nivel de correlación entre las medias y las varianzas. Es decir, pueden hacer las varianzas entre grupos más homogéneas.
- Muchas veces las transformaciones que afectan la distribución también mejoran problemas de homogeneidad de varianzas. De hecho y puesto que el supuesto de homogeneidad de varianzas es más crítico que el de normalidad, las transformaciones de datos se deben aplicar para solucionar el primer problema.

Clase 4: Tipos de Efectos, Modelo Lineal Simple de ANDEVA (CRD)

1. Tipos de “Tratamientos” o “Factores” en ANDEVA

En general hemos hablado de “tratamiento” al referirnos a la Fuente de Variación *entre grupos*. En forma más general, esta fuente de variación puede considerarse como un “**FACTOR**” con varios niveles (o grupos). Por ejemplo el *factor* “área de estudio” que usamos en uno de nuestros ejemplos tenía dos niveles: Estudiantes de biología y estudiantes de otras carreras. El factor Depredadores tenía también dos niveles: Con o sin la presencia de estos exudados. Un factor determinado puede tener varios *Niveles*. Por ejemplo el factor método de cultivo puede tener cuatro diferentes técnicas de cultivo, o el factor especie de árbol que recibe un determinado tratamiento puede tener cuatro niveles correspondientes a cuatro especies distintas.

En forma general existen DOS tipos diferentes de factores: FIJOS y ALEATORIOS. El nombre aleatorio NO tiene que ver con la aplicación de tratamientos en forma aleatoria o realizar un muestreo aleatorio. Tanto para factores fijo o aleatorios tenemos los mismos supuestos de ANDEVA.

La diferencia entre factores fijos y aleatorios radica principalmente en la interpretación los resultados en un análisis de varianza de una vía o de un factor y en la manera como los niveles de ese factor son incorporados al estudio.

Factores Fijos:

Un factor o fuente de variación se considera *fijo* si todos los niveles posibles de ese factor o al menos todos los niveles de interés para los investigadores son incluidos en el experimento o análisis

- Si un estudio quiere ver las diferencias entre sexos en las tasas de reabsorción de agua, debe considerar los dos sexos (machos y hembras) y estos son todos los sexos posibles. Entonces, el factor “sexo” es un factor fijo.
- Si un estudio que está interesado en ver el efecto de una hormona específica sobre las tasas de crecimiento de chanchos, a los investigadores les interesa solamente el efecto de esa hormona con respecto al control y entonces el factor es fijo.
- Si un estudio quiere ver el efecto de las estaciones del año sobre la producción de biomasa por plantas perennes, y el estudio considera las cuatro estaciones del año, entonces el factor “estación del año” es un factor fijo.

Una definición general para factores fijos es la siguiente:

Supongamos que existen N niveles posibles de un factor A ($A_1, A_2, A_3 \dots A_N$). De estos nosotros “muestreamos” o usamos a niveles en nuestro estudio o experimento. La fracción usada entonces es a/N . Entonces, $1-a/N = 0$ define un factor fijo.

Factores Fijos son descritos como Modelo I de ANDEVA

Factores Aleatorios:

Un factor o fuente de variación se considera *aleatorio* cuando los niveles de ese factor que son considerados en el experimento o estudio son elegidos en forma aleatoria de un universo mucho mayor de niveles posibles. En este caso, los niveles del factor que son de interés para los investigadores son más que aquellos directamente incorporados en el experimento.

- Si un estudio quiere determinar la relación general entre la temperatura y las tasas metabólicas y no existe ninguna razón *a priori* para seleccionar una temperatura en particular de todos los valores posibles de temperatura, entonces el factor “temperatura” es *aleatorio*.
- Si un estudio quiere investigar la diversidad genética espacial de una especie de ratón en la zona árida del país y toma muestras en cuatro localidades seleccionadas al azar dentro del rango de distribución del ratón, entonces el factor “localidad o sitio” es aleatorio.
- Si un estudio desea comparar la variación espacial en la intensidad de depredación por estrellas de mar a lo largo del país y realiza experimentos en seis sitios seleccionados al azar, entonces el factor “sitio” es aleatorio.
- Si un estudio quiere ver las interacciones entre la favorabilidad del ambiente (variables ambientales) y la expresión de dos genotipos distintos, y las variables ambientales son seleccionadas al azar de un rango de variables posibles, entonces el factor “ambiente” es un factor aleatorio.

Cuando los factores son aleatorios los niveles muestreados o usados en el estudio, a , son una fracción muy pequeña de los niveles posibles, N . Entonces $1-a/N \approx 1$.

Análisis de factores aleatorios son descritos como Modelo II de ANDEVA

DOS PREGUNTAS QUE AYUDAN A DETERMINAR SI UN FACTOR ES FIJO O ALEATORIO:

1. ¿Existe interés y/o se gana algo importante al realizar una prueba *a posteriori* y determinar que nivel o grupo difiere significativamente? Si es así, entonces el factor probablemente debe ser considerado como fijo

2. Si el experimento se repitiera nuevamente por uno mismo u otros investigadores, ¿se volverían a usar exactamente los mismos niveles del factor? Si es así, probablemente el factor debe ser considerado como fijo.

Cuando la fuente de variación entre tratamiento o “factor” de una ANDEVA es fijo, se dice que la ANDEVA es **Modelo I** (efectos o factores fijos).

Cuando la fuente de variación entre tratamientos o factor es aleatorio, se dice que la ANDEVA es **Modelo II** (efectos o factores aleatorios).

¿Existe alguna diferencia en los resultados de una ANDEVA simple entre Modelo I y Modelo II cuando el análisis incluye un solo factor o tratamiento (= *una vía*)?

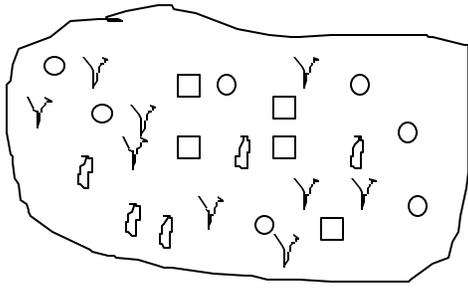
NO, el cálculo matemático y las pruebas de hipótesis de factores fijos y aleatorios son las mismas en el caso de ANDEVA simple (*de una vía*) de Modelo I y Modelo II.

Sin embargo, aún para el caso de ANDEVA simple de una vía, el procedimiento a seguir luego del análisis principal es diferente para ANDEVA Modelo I y Modelo II. Además, el cálculo de poder de la prueba de ANDEVA es diferente en el caso de Modelo I y Modelo II.

2. Modelo Lineal Simple para Diseño Completamente Aleatorio de un Factor

Supongamos que estamos interesados en saber si la tasa de fotosíntesis promedio de arbustos de la zona semiárida de una localidad en particular son significativamente diferentes entre especies.

Ahora supongamos, además, que existen solamente cinco especies de arbustos en esa localidad y nosotros consideramos todas las cinco especies en nuestro estudio, es decir, tenemos todas las especies posibles. Nuestra hipótesis entonces está referida a diferencias en tasas de fotosíntesis entre especies de arbustos de esa localidad y entonces el factor “especies de arbusto” es un factor fijo. Por supuesto lo que queremos hacer es seleccionar un número adecuado de arbustos de cada una de las especies (idealmente igual número para tener un diseño balanceado) y a cada arbusto medirle la tasa de fotosíntesis a una determinada hora del día.

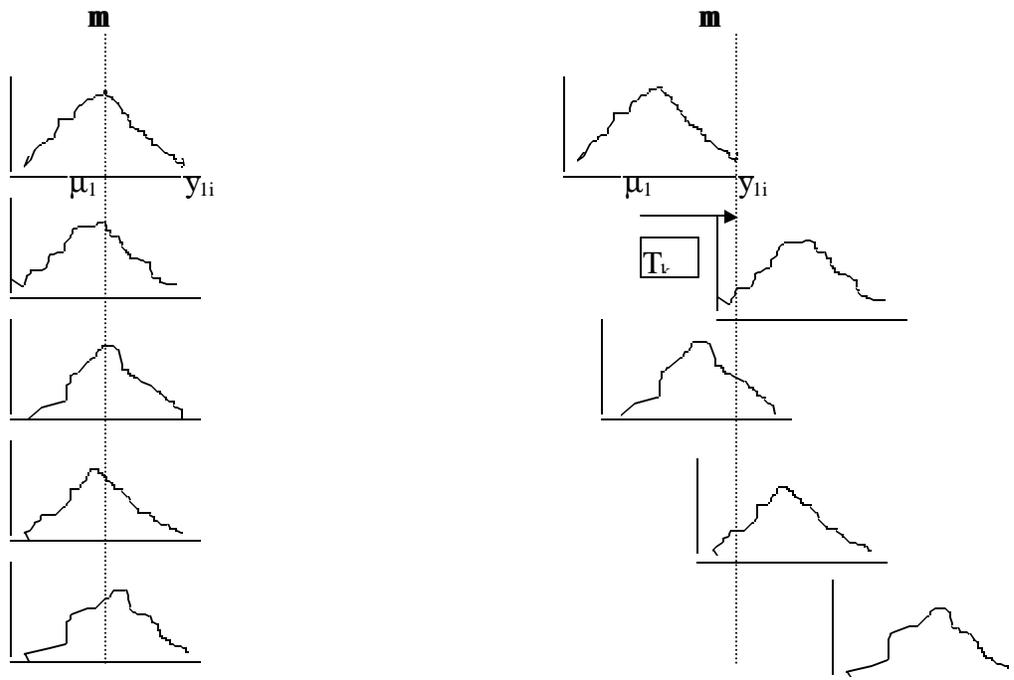


Los arbustos de cada una de las especies deben ser seleccionados al azar, por ejemplo a través de extender un transecto de 200 metros de largo con 10 números al azar para cada especie y muestreando al arbusto más cercano a cada uno de esos puntos. Tenemos entonces $a = 5$ y n (réplicas) = 10 para cada arbusto (balanceado).

La tasa de fotosíntesis de un arbusto determinado de la especie k estará dada por:

$$y_{ki} = \mu_k + e_{ki}$$

Donde μ_k es la media de tasa fotosintética para la especie k y e_{ki} es una medida de la varianza de la población de los individuos de la especie k . Este término mide cuan lejos esta la observación y_{ik} de la media de fotosíntesis para la especie $k=1$, o 2 , o 3 , etc. Ahora bien, en términos de la hipótesis nula, todas los grupos a comparar, en este caso las especies de arbustos, tienen la misma media, es decir, $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu$ (gran media)



Entonces, una observación determinada puede ser expresada como:

$$y_{ki} = \mu + T_k + e_{ki} \quad \text{Modelo Lineal para CRD}$$

En donde T_k se denomina el “efecto” del nivel o grupo k sobre la media de la población. La hipótesis nula entonces puede ser re-expresada como:

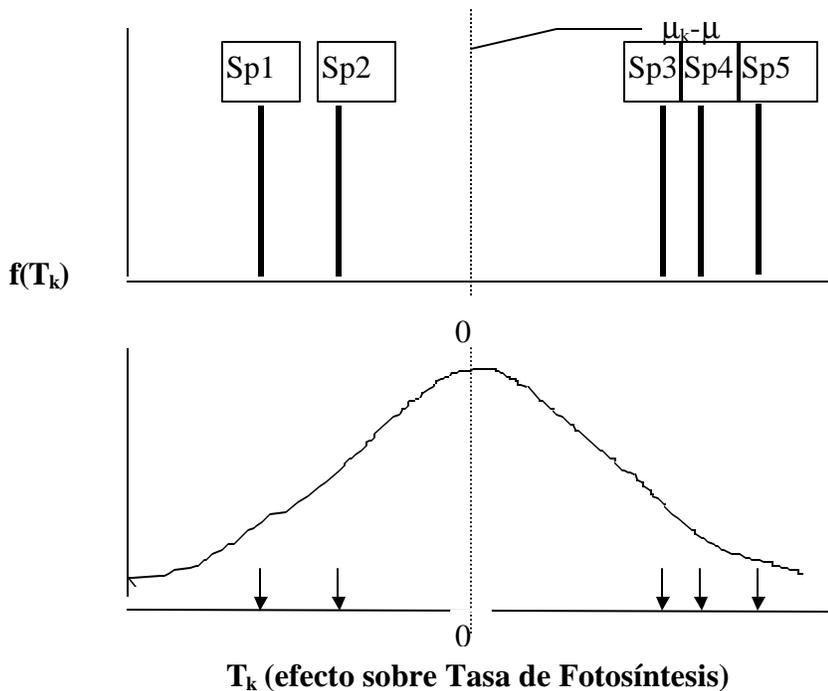
$$H_0: T_1 = T_2 = T_3 = T_4 = T_5 = 0$$

Además, se debe cumplir siempre que $\sum T_k = 0$. Es decir, que la sumatoria de efectos de todos los grupos debe ser cero. Este es un supuesto ANDEVA que no siempre se presenta en forma explícita. Se denomina “Supuesto de Sumatoria a Cero”. Violación del supuesto que la sumatoria de efectos es cero no afectará las pruebas de significancia en el ANDEVA. Sin embargo, si desea estimar la magnitud de los efectos de un tratamiento o variable sobre la variable respuesta, entonces se debe suponer que la sumatoria de todos los efectos es cero, o al menos que el valor esperado de esta sumatoria es cero (factores aleatorios).

Noten que el término e_{ik} sigue siendo la desviación de una observación de la media de cada tratamiento. La magnitud de estas desviaciones, de todas las observaciones es estimada por el error experimental o residuo.

3. Factores Fijos y Aleatorios: Magnitud de Efectos y Estimación de Varianzas Debida a un Factor

La diferencia entre factores fijos y aleatorios tiene una consecuencia obvia. En el ejemplo anterior, dijimos que estamos interesados en comparar las tasas de fotosíntesis de los cinco arbustos que habitan nuestra localidad de interés y para ello medimos las tasas fotosintéticas en 10 individuos representativos de cada una de las especies.



Cuando tenemos un factor fijo, entonces al calcular las sumas de cuadrados entre grupos o entre niveles, en realidad estamos calculando el “efecto” real que tienen las especies sobre las tasas fotosintéticas. Es decir, calculamos estas diferencias entre los factores de las especies, las que llamaremos T^2 .

Ahora bien, supongamos que en nuestra área de interés en realidad hay 40 especies de arbustos y no solamente cinco. Nosotros aun estamos interesados en saber si las tasas fotosintéticas son diferentes entre especies de arbustos, pero no podemos medirlas todas. Entonces decidimos “muestrear” o “seleccionar” al azar cinco especies de arbustos del total de especies presentes en el sitio. Estas especies son elegidas como representativas del efecto que el tratamiento “especie de arbusto” tiene sobre las tasas fotosintéticas, pero podríamos haber elegido otras especies. Ciertamente, si elegimos otras especies no tendremos exactamente los mismos efectos y por simple azar la sumatoria de efectos puede no ser igual a cero, lo que se espera es que en promedio, si repetimos el experimento varias veces con diferentes grupos de cinco especies entonces $\sum T_k = 0$

En este caso NO conocemos el efecto “real” o total del tratamiento “especie de arbusto”, sólo podemos estimarlo. La estimación es una estimación de la varianza de los efectos T o simplemente: S^2_T

4. Cuadrados Medios Esperados para ANDEVA de una vía Modelos I y II

Los Cuadrados Medios Esperados son la expresión de los componentes de varianzas que conforman los cuadrados medios (CM) en una tabla de ANDEVA. En otras palabras, cual(es) componentes de la varianza total se espera que nuestras “fuentes de variación” estén realmente estimando. Los cuadrados medios esperados no son valores numéricos, sino que ecuaciones que nos indican los componentes de varianza que conforman esas fuentes de variación.

Estas ecuaciones de cuadrados medios esperados tienen dos usos principales:

- a) Muestran como obtener estimaciones no sesgadas de los errores residuales que deben usarse en cada prueba de hipótesis de interés (denominador a usar en el calculo de F),
- b) Proveen una forma de estimar la contribución que realizan diferentes fuentes de variación a la varianza de la variable medida.

La forma de la expresión (ecuación) de cuadrados medios esperados, relacionada a cada fuente de variación, dependerá de si esa fuente de variación es un factor fijo o un factor aleatorio.

A. CME para el caso de ANDEVA de una vía Modelo I (factor fijo)

Tasas de fotosíntesis en las cinco especies de arbustos que habitan el área de interés (cuando tenemos todas las cinco especies del matorral).

Fuente de Variación	g.l	SC	CM	CME	F
Tratamiento	a-1		SCe/a-1	$\sigma_e^2 + nT^2$	$(\sigma_e^2 + nT^2) / \sigma_e^2$
Error	a(n-1)		SCd/a(n-1)	σ_e^2	
TOTAL	an-1				

En este caso:

1. El termino de Error o residual estima la varianza debida a las variaciones de cada arbusto dentro de cada especie. Esta es la varianza “no explicada” por la variable manipulada (factor o tratamiento)
2. El coeficiente de σ_e^2 es siempre 1
3. El cuadrado medio del factor o tratamiento (T) estima la varianza debida a la varianza dentro de tratamientos (σ_e^2), más la varianza debida al factor. Cuando el factor es fijo,
 $T^2 = (S (T_k - T)^2) / (a-1)$
4. El coeficiente de la varianza debida al factor será siempre n (replicación dentro de grupos o su respectivo n_i en caso de desbalance) por el producto del numero de niveles de todos los otros factores.

B. CME ANDEVA una-via Modelo II (factor aleatorio)

El estudio considera las tasas de fotosíntesis en cinco especies de arbustos elegidas al azar de las 40 especies que habitan el área de interés. La diferencia en este caso es que estamos interesados en las cuarenta especies y utilizamos una muestra de cinco de estas especies para estimar σ_T^2 , la varianza “producida” por pertenecer a las distintas especies de arbustos. En otras palabras, queremos saber cual es nivel de variabilidad *entre* especies de arbustos en relación con la variación entre individuos *dentro* de cada especie de arbusto. Es frecuente que en estudios de factores aleatorios estemos más interesados en conocer la contribución de estos factores a la varianza en la variable respuesta que en la significancia misma del factor.

Tabla de ANDEVA una vía Modelo II.

Fuente de Variación	g.l	SC	CM	CME	F
Tratamiento	a-1		SCe/a-1	$\sigma_e^2 + nS^2_T$	
Error	A(n-1)		SCd/a(n-1)	σ_e^2	
TOTAL	An-1				

La varianza o efecto que nos interesa estimar en ambos casos es T^2 o σ_T^2 y en el caso de un análisis de una vía la prueba de hipótesis es exactamente la misma: Dividimos los cuadrados medios entre grupos por los cuadrados medios dentro de grupos. Matemáticamente entonces, el considerar el factor como fijo o aleatorio en análisis de varianza de una vía no tiene efectos sobre el cálculo de Suma de Cuadrados y Cuadrados Medios, ni tampoco sobre la forma de realizar la prueba de hipótesis. Sin embargo, la interpretación de los resultados de estudios con distintos tipos de factores será, necesariamente, diferente.

Además, puesto que en un caso “conocemos” los efectos del factor de interés y en el otro “estimamos” la varianza producida por éste, los cálculos del Poder de ANDEVA Modelo I y Modelo II son diferentes.

Poder de ANDEVA cuando el factor es Fijo.

El cálculo de poder de una prueba de ANDEVA de una vía es relativamente simple. Como vimos anteriormente, el poder de una prueba está determinado por la tasa de Error Tipo I, la variabilidad observada en la variable respuesta (σ^2), el tamaño de los efectos ($\Sigma(T_i - T)^2$), y el tamaño muestral o replicación, n. En general, uno lo que hace es calcular el tamaño muestral necesario para obtener un determinado poder de rechazar una hipótesis nula falsa. Para ello se debe primero determinar la magnitud de los efectos que se considera importante a la luz de nuestra hipótesis alternativa.

En el caso de un factor fijo podemos expresar el cálculo de F como:

$$F = \frac{CM_{entre}}{CM_{dentro}} = \frac{s_e^2 + \frac{n \sum_{k=1}^a (T_k - \bar{T})^2}{a-1}}{s_e^2} = 1 + \frac{1}{a-1} n \frac{\sum_{k=1}^a (T_k - \bar{T})^2}{s_e^2}$$

donde,

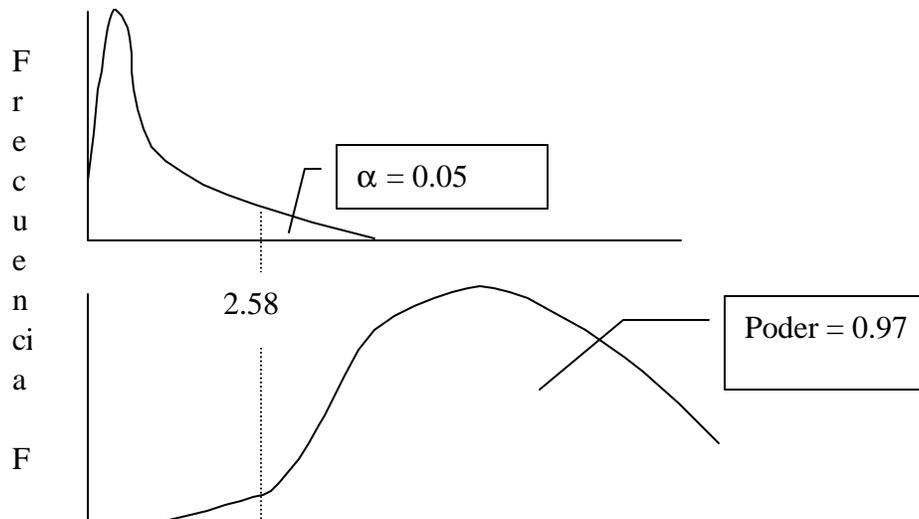
$$I = \frac{n \sum_{k=1}^a (T_k - \bar{T})^2}{s_e^2}$$

el término λ se denomina *parámetro no central*. Con este parámetro no central se calcula la función

$$f = \sqrt{\frac{I}{a}} = \sqrt{\frac{n \sum_{k=1}^a (T_k - \bar{T})^2}{a s_e^2}}$$

Cuando la hipótesis nula es verdadera, entonces todos los T_k son cero, la $\sum T^2 = 0$ y el valor esperado de F es uno. Es decir, en un experimento determinado cuando H_0 es verdadera, el valor de F proviene de la *distribución central* de F que vimos anteriormente, con $\nu_1 = a-1$ y $\nu_2 = a(n-1)$ grados de libertad.

Si por el contrario la hipótesis nula es falsa, $\sum T^2 > 0$, significa que existe una determinada hipótesis alternativa que es correcta (que probablemente sea nuestra hipótesis alternativa!) y que especifica determinados valores de magnitudes de efectos de cada grupo (T_k). Bajo estas circunstancias, el valor de F ya no proviene de la distribución central especificada por la hipótesis nula, sino que de la *distribución de F no central* especificada por esa hipótesis alternativa y con los mismos grados de libertad (ν_1, ν_2).



Valor de F

El calculo de poder depende críticamente de la determinación, antes de realizar un estudio determinado, de la magnitud de las diferencias mínimas entre los grupos a comparara que apoya nuestra hipótesis alternativa.

Supongamos que deseamos calcular el poder de una prueba de ANDEVA que realizaremos para comparar los efectos de exudados de tres especies de depredadores sobre el grosor de la concha de choritos intermareales. Los depredadores seleccionados son el loco, el sol de mar y jaibas depredadoras. Para poder evaluar el efecto de los tratamientos, el estudio incluye además un control sin depredador. Si estas son las tres especies de depredadores de interés, entonces el factor 'especie de depredador' debe ser fijo. Nuestra hipótesis alternativa es la motivación para realizar el estudio y nos debe dar una indicación de las diferencias esperadas. Supongamos que nuestra hipótesis predice que de las tres especies de depredadores, solamente una de las especies induce mayores grosores de concha, esta es la jaiba que necesita quebrar la concha para consumir choritos. Los otros depredadores con quiebran la concha para consumir los choritos. Lamentablemente, sólo indicar cuál especie esperamos que difiera significativamente no es suficiente para calcular el poder de una prueba, pues necesitamos calcular el parámetro no central y ϕ . Supongamos que nuestro estudio esta efectivamente motivado por observaciones de variabilidad en el grosor de la concha de choritos. Choritos provenientes de lugares con altas densidades de jaibas tienen conchas de alrededor de 600 micras milímetros, mientras que choritos de zonas con pocas jaibas tienen conchas de alrededor de 400 micras. Si las jaibas son responsables de esta variación, entonces esperamos que los choritos en presencia de exudados de jaibas tengan conchas 200 micras mas gruesas que todos los otros. Esto es:

$$T_{\text{control}} = \text{Efecto control sin depredadores} = 0$$

$$T_{\text{soles}} = \text{Efecto soles} = 0$$

$$T_{\text{locos}} = \text{Efecto locos} = 0$$

$$T_{\text{jaibas}} = \text{Efecto de jaibas} = + 200$$

Para cumplir con el supuesto de sumatoria de efectos igual a cero ($\sum T_k = 0$), entonces los efectos serán: $T_c = -50$, $T_s = -50$, $T_l = -50$, $T_j = +150$. (La diferencia entre los efectos es de 200).

$$\text{Entonces, } \sum (T_k - T)^2 = 30000$$

SI tenemos una estimación de variabilidad en la variable respuesta, grosores de concha, entonces podemos calcular el parámetro no central. Es probable que si tenemos las

observaciones que motivaron el estudio también podemos calcular de allí una varianza. Supongamos que esta varianza entre las réplicas es $\sigma^2 = 5000$ (micras cuadradas). Con estos datos podemos calcular el poder de nuestra prueba, o mucho mejor, calcular el tamaño muestral (replicación) de nuestro experimento para obtener un poder determinado.

Lo primero es calcular ϕ :

$$f = \sqrt{\frac{I}{a}} = \sqrt{\frac{n \sum_{k=1}^a (T_i - \bar{T})^2}{a S_e^2}} = \sqrt{\frac{n * 30000}{4 * 5000}}$$

Con este valor de ϕ , y con un determinado nivel de error Tipo I predeterminado (normalmente 0.05) consultamos una tabla de *Distribución no central de F*, usando $\nu_1 = a - 1 = 3$, y $\nu_2 = 4(n - 1)$ grados de libertad. Estas tablas han sido tabuladas pensando principalmente que uno desea determinar el nivel de replicación, n , para obtener un determinado poder $(1 - \beta)$. (ver ejemplo de la estructura de estas tabla más abajo). Sin embargo, en la mayoría de las tablas el nivel de replicación adecuado para obtener un determinado poder debe calcularse por “prueba y error” (iteración), es decir probando valores de replicación iniciales y subiendo o bajando el n hasta lograr el poder deseado.

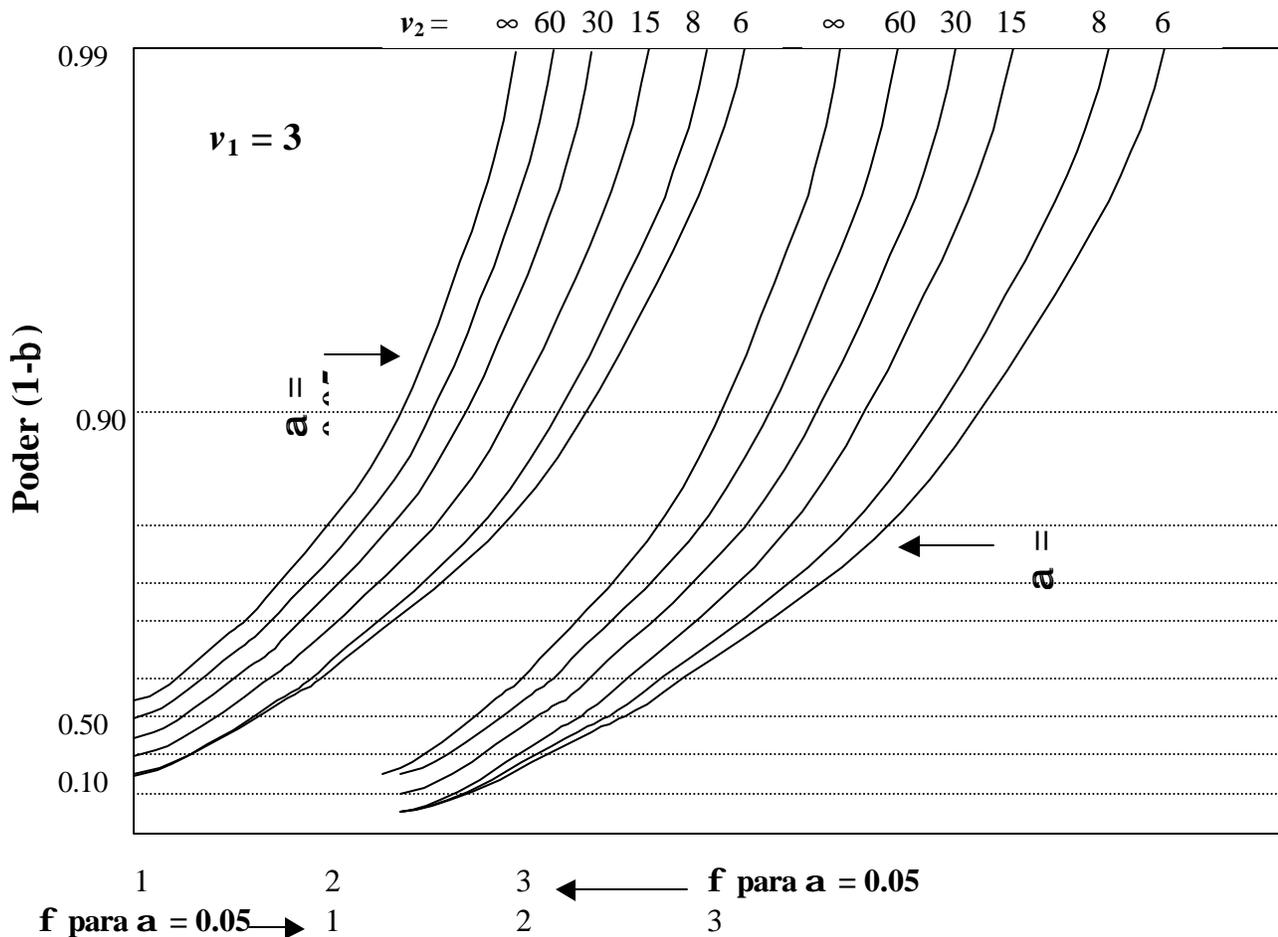
Por ejemplo, si desea someter a prueba la hipótesis de los efectos de depredadores usando una probabilidad de error Tipo I, $\alpha = 0.05$ y mantener una probabilidad de rechazar la hipótesis H_0 si esta es falsa de al menos 0.95 (Poder = 95%), entonces podemos empezar a probar con un $n = 4$. Esto nos dará un $\phi = 2.45$ y $\nu_2 = 4(4 - 1) = 12$. Usando la tabla de distribución F para curvas de poder de un factor fijo obtenemos un poder de entre 93 y 94%. Entonces, debemos aumentar un poco el n a 5. En este caso nos da un valor $\phi = 2.74$ y $\nu_2 = 4(5 - 1) = 16$. Usando la Tabla de poder de F encontramos que esta combinación nos dará un poder de más del 98%, por lo que este nivel de replicación parece suficiente .

Las tablas de distribución no central de F se encuentran en varios libros más o menos especializados (e.j. Cohen 1977, Kuhel, 1994 o Winer et al. 1984). En eneral los textos básicos no traen estas tablas. Winer et al. han generado tablas de “poder contante” que se pueden usar para calcular el nivel de replicación necesario para obtener un determinado poder, sin necesidad de probar valores de n . En esas tablas uno puede leer directamente el nivel de replicacion. Para ello, es necesario calcular una forma modificada de la función ϕ' (phi prima). Este valor ϕ' se calcula igual que ϕ pero sin considerar n . Winer et al. solamente entregan valores tabulados para un set limitado de valores de poder y no hay pocos textos que incluyan estas tablas, de manera

que nosotros solamente usaremos las tablas más tradicionales de poder y jugaremos “probando” niveles de replicación para obtener el poder deseado.

Cohen, J. 1977. *Statistical power analysis for the behavioral sciences*. Academic Press, New York.

Ejemplo de la **Tabla de Curvas de Poder** para prueba de F para factor fijo. Existe una tabla distinta para cada valor de ν_1 (grados de libertad del numerador) o número de grupos a comparar. El nivel de Error Tipo I aceptable, o significancia (α) determina el grupo de curvas que deben seguirse y cual es la línea de valores del parámetro no central estandarizado (ϕ) que debe leerse (en la base de la Tabla). El nivel de replicación determinará los grados de libertad del denominador (ν_2) y con ello la curva de poder específica a seguir. Si se desea encontrar el nivel de replicación adecuado para obtener un determinado poder (ej. 90%), se debe empezar a “probar” valores de n hasta obtener el nivel de replicación que nos de el 90% de poder.



Clase 5: Diseños con Submuestras, Ejemplos

1. Diseño Completamente Aleatorio con Submuestras (= Diseño Anidado)

Ahora vamos a hacer el mismo ejemplo de tasas de fotosíntesis de arbustos pero un poco más realista. La verdad es que uno no puede simplemente medir la tasa de fotosíntesis de un arbusto en una simple medición, aun y cuando uno especifique la hora. Existen dos problemas:

Primero, diferentes partes de los arbustos presentarán diferentes tasas de fotosíntesis. Por ejemplo las hojas e la parte superior del follaje presentarán tasas más altas que las del interior del follaje o el tallo de las hojas será diferente que la parte apical de la hoja, etc. Para resolver este problema tenemos varias posibilidades:

a) Modificar nuestra hipótesis para especificar que parte del arbusto compararemos (diferencias entre las hojas apicales del follaje superior de las distintas especies de arbustos) y entonces realizar mediciones en estas áreas solamente.

b) La otra posibilidad es incorporar esta variación en nuestra hipótesis y en el diseño de muestreo. Esto lo veremos un poco más adelante en diseños anidados.

El otro problema es que aún cuando decidimos medir las tasas de fotosíntesis en una parte del arbusto determinada, como la zona apical de las hojas del follaje superior, aun existe una gran variación entre las hojas de esta misma zona. Para obtener una mejor estimación de las tasas de fotosíntesis de esta zona de cada arbusto, decidimos entonces medir la fotosíntesis en 5 hojas de cada arbusto.

recuerden que aún tenemos 10 arbustos por especie, pero ahora además tenemos una medida de variación entre cinco hojas de un mismo arbusto.

Este modelo puede expresarse así:

$$y_{kij} = \mu + T_k + J_{ij} + e_{kij}$$

En este caso, la pregunta más importante que uno debe hacerse es ¿Cuales son nuestras réplicas? Los valores promedios de los diez individuos o los valores de cada una de las hojas de cada individuo?

Tenemos $a = 5$ especies de arbustos

Tenemos $n = 10$ arbustos de cada especie

Tenemos $s = 5$ mediciones de fotosíntesis en 5 hojas

Fuente de Variación	g.l	SC	CM	CME	F
Tratamiento	a-1 = 4		SCE/a-1	$\sigma_h^2 + s\sigma_a^2 + snS^2_T$	
Error debido a arbustos	a(n-1) =45		SCa/a(n-1)	$\sigma_h^2 + s\sigma_a^2$	
Error debido a hojas	an(s-1) = 200		SCh/an(s-1)	σ_h^2	
TOTAL	ans-1= 249				

Entonces: ¿Cuál es el error apropiado para someter a prueba la hipótesis de no diferencias entre las tasas de fotosíntesis de diferentes especies?

Claramente el error experimental para esta hipótesis es el error debido a las diferencias entre arbustos y NO el error debido a diferencias entre hojas dentro de arbustos.

Las diferencias entre hojas dentro de un mismo arbusto no son medidas independientes entre si para estimar la tasa de fotosíntesis de la población de esa especie en el área. Estas tasas de fotosíntesis estarán más correlacionadas entre si puesto que pertenecen a un mismo individuo.

Cuando usamos incorrectamente la fuente de variación debido a las diferencias entre las hojas, las cuales tienen 200 grados de libertad asociados a la estimación de varianza, en vez de la variación debida arbustos con 45 grados de libertad, estamos cometiendo o realizando pseudoreplicación. Hurlbert en 1984 escribió una monografía completa dedicada a este problema de pseudoreplicación y cuan común es entre ecólogos usar términos de error equivocados para someter a prueba hipótesis.

El diseño anterior también nos dará una prueba de hipótesis adecuada si antes de realizar el análisis calculamos el promedio de las cinco hojas para cada uno de los arbustos. Entonces el modelo se reduce al modelo CRD simple.

La ventaja de incluir esta fuente de variación en el modelo es que ahora podemos además realizar una prueba de hipótesis para ver si los arbustos de una misma especie varían entre si y podemos saber la magnitud de la variación de las hojas dentro de un mismo arbusto. Esto lo veremos en mayor detalle cuando veamos diseños anidados.

Los diseños con submuestras son muy comunes en todas las ramas de la biología. Con frecuencia investigadores deben realizar varias mediciones de un mismo individuo, tomar varias

lecturas de concentración en un espectrofotómetro o correr varios geles de una misma muestra sólo para corroborar la existencia de bandas. No existe ningún problema con estos diseños, excepto que los investigadores deben reconocer cual es la fuente de error experimental, es decir, cuáles son las unidades experimentales que pueden considerarse independientes. La consideración adecuada de las unidades experimentales no es solamente importante para realizar prueba de hipótesis con la fuente de error dentro de tratamientos apropiada, sino que también para calcular el Error Estándar apropiado cuando se presentan figuras o tablas.

Es muy común que los investigadores calculen los errores estándar que se presentan en figuras y tablas usando todas las observaciones bajo un determinado tratamiento ($n * s$). Esto por supuesto reduce enormemente y erróneamente la magnitud de las barras de error en las figuras. El cálculo de intervalos de confianza, depende de la estimación correcta del error estándar.

II. EJEMPLO DE DISEÑOS COMPLETAMENTE ALEATORIOS

Un biólogo es contratado para realizar un estudio acerca del efecto que pueden tener los efluentes de una industria sobre los peces que habitan un río sobre el cual la industria vierte sus efluentes. Parte de los efluentes contienen concentraciones elevadas de algunos metales pesados y eso lleva al biólogo a pensar que los efluentes pueden tener un efecto sobre las tasas de crecimiento de los peces.

Como una primera aproximación al problema, el biólogo diseña un experimento de laboratorio para comprobar si los efluentes afectan las tasas de crecimiento de una especie de pez nativa del río y la más abundante. El experimento consiste en exponer a peces a aguas del río sin efluentes y a aguas del río con efluentes de la planta.

Las hipótesis del biólogo es:

Ho: Los efluentes de la planta no tienen efectos sobre las tasas de crecimiento de los peces de la especie *Statisticus nosabemus*

Ho: $\mu_1 = \mu_2$

Ha: Los efluentes tienen un efecto significativo sobre las tasas de crecimiento de los peces.

Ha: $\mu_1 \neq \mu_2$

- Para realizar el experimento el biólogo cuenta con 10 acuarios y sabe que, si no existen razones a priori que indiquen lo contrario, el mejor diseño para este experimento es un diseño completamente aleatorio y un análisis de varianza de una vía.
- Además, sabe que si tiene 10 acuarios y dos tratamientos, lo mejor es designar, en forma aleatoria 5 acuarios a un tratamiento y 5 al otro para tener un diseño balanceado.

- Puesto que los peces normalmente se agrupan en cardúmenes y también para aumentar la precisión de las medidas de crecimiento, el biólogo decide poner 6 peces más o menos del mismo tamaño en cada acuario y en forma aleatoria.
- Al cabo de 3 semanas el biólogo mide cada individuo de cada acuario y ahora quiere analizar los datos.

a = 2 tratamientos

n = 5 acuarios

s = 6 peces por acuario

DATOS ORIGINALES PARA CADA INDIVIDUO MEDIDO EN EL EXPERIMENTO			
a	i	j	W
Trat	Acuario	Indiv	Peso
1	1	1	101.468
1	1	2	103.776
1	1	3	95.122
1	1	4	125.098
1	1	5	105.685
1	1	6	98.417
1	2	1	109.05
1	2	2	97.013
1	2	3	106.538
1	2	4	97.377
1	2	5	104.295
1	2	6	96.936
1	3	1	98.889
1	3	2	105.275
1	3	3	108.92
1	3	4	111.597
1	3	5	109.639
1	3	6	84.823
1	4	1	90.139
1	4	2	114.182
1	4	3	98.458
1	4	4	95.112
1	4	5	96.486
1	4	6	94.309
1	5	1	110.683
1	5	2	106.128
1	5	3	85.182
1	5	4	117.069
1	5	5	111.868
1	5	6	87.279
2	1	1	75.396

ESTADISTICOS POR ACUARIO (REPLICA)					
a	i	W			n
Trat	Acuario	Promedio	Varianza	EE	
1	1	104.93	111.85	4.32	6
1	2	101.87	29.47	2.22	6
1	3	103.19	101.06	4.10	6
1	4	98.11	69.61	3.41	6
1	5	103.04	182.01	5.51	6
2	1	86.84	107.31	4.23	6
2	2	84.99	31.09	2.28	6
2	3	81.87	35.45	2.43	6
2	4	88.71	83.04	3.72	6
2	5	86.67	102.69	4.14	6

ESTADISTICOS POR TRATAMIENTO				
a	W		EE	n
Trat	Prome	VAR		
	dio			
1	102.22	6.478	1.13832	5
2	85.814	6.599	1.14889	5

2	1	2	83.956
2	1	3	94.6
2	1	4	78.536
2	1	5	85.397
2	1	6	103.149
2	2	1	91.479
2	2	2	86.801
2	2	3	90.461
2	2	4	80.049
2	2	5	77.69
2	2	6	83.434
2	3	1	88.858
2	3	2	77.089
2	3	3	74.677
2	3	4	82.764
2	3	5	88.639
2	3	6	79.187
2	4	1	79.326
2	4	2	103.093
2	4	3	86.144
2	4	4	95.239
2	4	5	88.141
2	4	6	80.3
2	5	1	97.916
2	5	2	92.023
2	5	3	95.233
2	5	4	71.274
2	5	5	82.086
2	5	6	81.4787

- Lo primero que debe hacer es explicitar las hipótesis. Eso ya lo hicimos.
- Luego, al definir las hipótesis debe decidir si el factor bajo estudio es fijo o aleatorio. En este caso el factor es fijo (con o sin efluentes)
- Luego debe decidir qué diseño debe usar y cuáles son las replicas del experimento. En este caso el biólogo piensa que cada acuario es una réplica independiente, puesto que los seis peces están sometidos al mismo ambiente y pueden afectarse unos a otros dentro de cada acuario.
- Ahora, también debe decidir que nivel de significancia usará para tomar la decisión de aceptar o rechazar la hipótesis nula. En este caso, seguirá el valor convencional de 0.05.
- Ahora que ya tiene los datos debe asegurarse que los datos sean “normales”.

¿A qué datos debe comprobar la normalidad? ¿A los acuarios o a los individuos?

- Para someter a prueba la hipótesis de diferencias entre los tratamientos, debe verificar que los promedios de los acuarios provienen de una distribución normal. Para esto contará con 5 puntos para realizar una gráfica de cada tratamiento.
- También puede verificar si las observaciones en cada uno de los acuarios sigue una distribución normal puesto que so lo son, los promedios de estas distribuciones también deben seguir una distribución normal.

- Alternativamente y aún mejor, puede verificar el supuesto de normalidad sobre los residuos luego de ajustar el modelo adecuado.

Supongamos que el biólogo no sabe como trabajar con un modelo con jerarquía y decide simplemente obtener el promedio de cada acuario (sobre los seis peces por acuario) y sobre estos promedios realizar el análisis. El modelo adecuado para este análisis es:

$$y_{ki} = \mathbf{m} + \mathbf{T}_k + \mathbf{e}_{ki}$$

Aquí esta la gráfica de los residuos de este modelo, la cual se ve relativamente normal. El gráfico Q-Q que muestra que los datos caen más o menos sobre la línea recta y el test de Kolmogorv-Smirnov para verificar el grado de ajuste entre las frecuencias observadas y aquellas esperadas bajo el supuesto de normalidad.

Ahora necesitamos verificar el supuesto de **homogeneidad de varianzas**. La prueba C de Cochran aplicada a los datos muestra que no hay diferencias significativas entre las varianzas de los grupos.

- Ahora el biólogo realiza el análisis. Tiene sus datos en la forma que se muestra en el output y usando un paquete estadístico, pide realizar un ANDEVA de una vía sobre los datos originales.

El output de este análisis se muestra en la Tabla de ANDEVA 1.

Tabla 1. ANDEVA para CRD una-via sobre los acuarios (intención)

Dependent Variable: W

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Tratamiento	1	4040.90703	4040.90703	51.15	0.0001
Error	58	4581.70916	78.99499		
Corrected Total	59	8622.61619			

¿Qué pueden decir de este análisis?

El programa, como la mayoría de los programas estadísticos ha asumido que cada una de las observaciones, basadas en los individuos es una observación independiente, es decir una réplica y por ello ahora tenemos **a((n s) -1)**, lo cual obviamente es incorrecto.

El análisis correcto de este diseño en el cual NO incorporamos la variación entre individuos en un mismo acuario es realizar el análisis sobre el promedio de los acuarios.

Tabla 2. ANDEVA para CRD de una via sobre los acuarios (correcto)

Dependent Variable: W

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Tratamiento	1	673.484504	673.484504	102.99	0.0001
Error	8	52.313996	6.539249		
Corrected Total	9	725.798500			

- Ahora concluimos que existen diferencias significativas sobre entre los grupos y basados en las medias podemos decir que los peces sometidos a los efluentes de la planta crecen significativamente menos que los controles.

Digamos ahora que el biólogo en cuestión tomó esta clase de estadística y sabe que uno puede usar la información de los individuos para someter a prueba otra hipótesis adicional, que se refiere a si los acuarios entre si, dentro de un mismo tratamiento, difieren significativamente. Esto es importante puesto que podemos ver a) si grupos de peces, o cardúmenes en particular se comportan de manera parecida y b) podemos realizar un análisis de componentes de varianza y ver si cuando la industria quiera repetir los experimentos deberíamos invertir más en aumentar el número de individuos por acuario o aumentar el número de acuarios por tratamiento. Esto dependerá de cuan variables son los individuos dentro de un mismo acuario.

Entonces, el modelo completo con jerarquía es:

$$y_{kil} = \mathbf{m} + \mathbf{T}_k + \mathbf{J}_{ki} + \mathbf{e}_{kij}$$

Ho2: No existen diferencias entre los acuarios dentro de tratamientos (todos los cardúmenes de peces se comportan igual)

Ha2: Cada cardumen de peces reacciona en forma diferente a la presencia o ausencia de efluentes (existen diferencias significativas entre acuarios dentro de tratamientos)

El biólogo realiza el análisis y el programa entrega los resultados de la Tabla 3.

Tabla 3. ANDEVA una via con jerarquía.

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Tratamiento	1	4040.90703	4040.90703	47.34	0.0001
Acuarios(trat)	8	313.88397	39.23550	0.46	0.8784
Error	50	4267.82519	85.35650		
Corrected Total	59	8622.61619			

En este caso ¿Cuál es el problema?

El programa NO sabe cuales son las varianzas esperadas pues esto dependen de las particularidades del diseño y de si nuestros efectos son fijos o aleatorios. Si interpretamos estos resultados estaremos **pseudoreplicando!**

En nuestro caso, dijimos que el tratamiento es un efecto FIJO. Ahora bien, cada cardumen asignado a un acuario en particular es DEBE ser un efecto ALEATORIO, puesto que se considera que este nivel es una réplica de nuestro tratamiento. Por ello, uno NO puede elegir un grupo de peces en particular por alguna razón específica. Si ello es así, entonces debemos incluir esto como un tratamiento en nuestro diseño. Este tipo de diseño con más de un tratamiento es un diseño factorial y lo veremos más adelante.

Entonces, Los Cuadrados Medios Esperados para este diseño son:

Fuente de Variación	g.l.	CME	
Tratamiento	a-1	$\sigma_e^2 + \sigma_a^2 + n T^2$	$F = CM_T / CM_a$
Acuarios(trat)	a(n-1)	$\sigma_e^2 + s \sigma_a^2$	$F = CM_a / CM_e$
Error	an(s-1)	σ_e^2	
TOTAL	ans-1		

Si dividimos entonces la estimación de varianza (Cuadrados medios) debido a los tratamientos por la estimación de varianza (Cuadrados medios) debida a las diferencias entre acuarios tendremos:

$$F = 4040.9/39.23 = 102.98$$

Es decir, el mismo valor que si realizamos el análisis sobre los promedios de cada acuario, con la ventaja que ahora además tenemos una idea de las varianzas entre acuarios y si estos difieren significativamente o no.

Clase 6: Prueba t de Student, Pruebas de Distribución Libre, Pruebas *a posteriori*

I. RECAPITULACION

En la clase anterior vimos un ejemplo de un estudio realizado por un biólogo que trabaja para una consultora ambiental y al cual se le ha encargado un estudio del efecto de los efluentes de una planta sobre una especie nativa de pez, *Statisticus nosabemus*, que habita el río en donde la planta vierte los efluentes.

El biólogo ha diseñado un estudio experimental de laboratorio como una aproximación al problema y decide exponer peces a agua del río con y sin efluentes de la planta. Dispone de 10 acuarios en total y en forma completamente aleatoria asigna uno de los tratamientos (con o sin efluentes) a cinco acuarios y el otro a los otros cinco acuarios o unidades experimentales. También dijimos que para obtener un mayor realismo biológico y para mejorar la precisión de las mediciones de crecimiento, decide poner 6 peces en cada acuario.

El diseño es completamente aleatorio, con la modificación de que existen submuestras tomadas dentro de cada una de las unidades experimentales independientes.

a = 2 tratamientos

n = 5 acuarios

s = 6 peces por acuario

Entonces, el modelo completo con jerarquía debido a las submuestras es:

$$y_{kij} = \mathbf{m} + \mathbf{T}_k + \mathbf{J}_{ki} + \mathbf{e}_{kij}$$

Analizamos este experimento de dos maneras diferentes. Primero, sin considerar en forma explícita que en cada acuario tenemos 6 peces, es decir reduciendo nuestro modelo a $y_{ki} = \mathbf{m} + \mathbf{T}_k + \mathbf{e}_{ki}$, y segundo usando el diseño expandido.

Este último modelo se llama simplemente CRD

EL modelo con jerarquía o submuestras es caso de diseño ANIDADO

Con esto mostramos los resultados del análisis estadístico de estos datos y mostramos porque es importante, primero que nada, calcular manualmente cuales son o deben ser

los grados de libertad de nuestro diseño. Si no lo hacemos, el programa puede entregarnos resultados equivocados y no nos daremos cuenta. En segundo lugar también debemos calcular cuales deben ser los Cuadrados Medios Esperados, pues solamente al hacer esto sabremos cual es la fuente de error apropiada para someter a prueba la hipótesis de efecto significativo de los tratamientos. En este caso dijimos que la fuente de error apropiada para los tratamientos la varianza debida a las diferencias entre acuarios y no aquella debida a las diferencias entre individuos dentro de acuarios, por lo que debemos dividir el Cuadrado Medio de los tratamientos pro los Cuadrados Medios de los acuarios para obtener el F y no por los cuadrados medios del error debido a las diferencias entre individuos dentro de tratamientos.

Tabla 3. ANDEVA una vía con jerarquía.

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Tratamiento	1	4040.90703	4040.90703	47.34	0.0001
Acuarios(trat)	8	313.88397	39.23550	0.46	0.8784
Error	50	4267.82519	85.35650		
Corrected Total	59	8622.61619			

El resultado correcto de este análisis entonces es $F = 4040.9/85.37 = 102.99$

Ahora vamos a ver otra solución paramétrica para este análisis. La famosa prueba de t de Student.

II. PRUEBA DE T DE STUDENT PARA DIFERENCIAS ENTRE DOS GRUPOS

1. *Introducción*

La prueba de t de Student es una de las pruebas de hipótesis estadísticas más tradicionales en biología. En muchas áreas de la biología es prácticamente la única prueba estadística utilizada, muchas veces incorrectamente.

La principal razón de su popularidad (y abuso) en algunos campos de la biología, como en el área molecular, es la simplicidad de cálculo. Una prueba de t-Student se puede realizar fácilmente a mano o en una planilla de datos tipo Excel, con la gran ventaja de no tener que aprender nuevo software computacional. Hoy en día, con la accesibilidad a programas de estadística fáciles de usar (con menús e interfaces gráficas) resulta tan fácil el realizar una ANDEVA como una prueba de t de Student.

2. Prueba de Hipótesis en Prueba de t-Student:

Las prueba de hipótesis de una prueba de t se refiere a la igualdad de las medias de dos poblaciones, al igual que en un análisis de varianza. A diferencia de la ANDEVA, en la prueba de t solamente se pueden comparar dos grupos.

Ho: $m_1 = m_2$ y Ha: $m_1 \neq m_2$ (dos colas)

Ho: $m_1 = m_2$ y Ha: $m_1 > m_2$ (una cola)

- Al contrario de Análisis de Varianza, una prueba de t de Student puede ser de una sola cola.

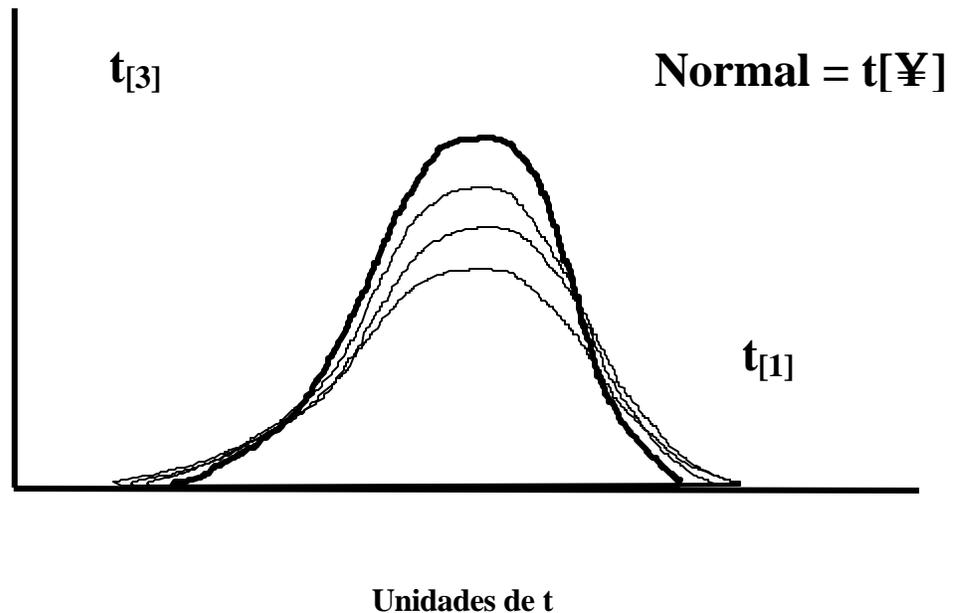
3. La idea básica del test de t de Student

La expresión:

$$\frac{(\bar{Y} - m)}{s_y}$$

sigue una distribución de t, en donde el numerador es la diferencia entre la media muestral y la media paramétrica y el denominador es la desviación estándar de esa desviación, es decir el Error Estándar.

Distribución de t-Student



El número de grados de libertad de la distribución de t es igual al número de grados de libertad de la desviación estándar en la razón $(m-\mu)/(s/\sqrt{n})$. A medida

que los grados de libertad de la distribución de t aumentan, la forma se hace más cercana a la distribución normal. La distribución de t con 30 grados de libertad es indistinguible de la distribución normal.

Los valores de la distribución de t se encuentran tabulados en tablas especiales. El uso de estas tablas se vio en el curso BIO-242 y lo verán en las sesiones prácticas.

Así también, la siguiente expresión:

$$t_s = \frac{(\bar{Y}_1 - \bar{Y}_2) - (\mathbf{m}_1 - \mathbf{m}_2)}{\sqrt{\left[\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \right] \left(\frac{n_1 + n_2}{n_1 n_2} \right)}}$$

sigue una distribución de t con $n_1 + n_2 - 2$ grados de libertad.

El numerador es una desviación entre la diferencia entre dos medias muestrales (Y barras) y la diferencia real existente entre las dos medias poblacionales.

En nuestra hipótesis nula estas dos muestras provienen de la misma población. Es decir, la diferencia entre las medias poblacionales se supone que es cero.

El denominador de la expresión es el Error Estándar de la diferencia entre dos medias.

Esto es, la raíz cuadrada de las varianzas, pesadas por las diferencias de tamaños muestrales entre los dos grupos (diseño desbalanceado) y divididas por los grados de libertad. Esto es el promedio de las *varianzas dentro de los grupos*.

El promedio de las varianzas dentro de los dos grupos debe ser multiplicado por $1/n_1 + 1/n_2$ ($= n_1 + n_2 / n_1 n_2$) para convertirlo en la *varianza de las diferencias entre los dos grupos*.

La semejanza con el estadístico calculado en ANDEVA es obvia. La diferencia entre las medias de los grupos es proporcional a las varianzas entre grupos y el error estándar promedio dentro de los grupos es una medida de varianza dentro de grupos.

Cuando los tamaños muestrales son iguales (diseño balanceado), entonces esta ecuación se reduce a:

$$t_s = \frac{(\bar{Y}_1 - \bar{Y}_2) - (\mathbf{m}_1 - \mathbf{m}_2)}{\sqrt{\frac{1}{n}(s_1^2 + s_2^2)}}$$

Esta expresión sigue una distribución de t con **2(n-1)** grados de libertad.

Para el ejemplo anterior: Experimento de efectos de efluentes sobre el crecimiento de peces.

Primero que nada, la prueba de t NO puede considerar los diseños jerárquicos, de manera que el único modelo posible es el modelo lineal simple: $y = \mu + T + e$

De esta manera, DEBEMOS calcular el promedio de todos los individuos en cada acuario antes de realizar la prueba de t. Usando estos valores y reemplazando en la ecuación,

$$\text{tenemos } t_s = \frac{(102.2 - 85.8) - (\mathbf{m}_1 - \mathbf{m}_2)}{\sqrt{\frac{1}{5}(6.48 + 6.59)}} = 16.4 / 1.62 = 10.12$$

Ahora debemos buscar el valor crítico de t para 0.05, de DOS COLAS, con 8 grados de libertad: Este valor es

$t_{0.05[8]} = 2.306$, mucho menor que el valor observado y entonces debemos rechazar la H_0 al nivel de significancia del 5%

La semejanza entre la prueba de t de Student y la ANDEVA no es sólo aparente o estructural. En realidad son matemáticamente equivalentes. De hecho t^2 es igual a F con 1 y ν grados de libertad:

$(10.12)^2 = 100.5$ casi exactamente el valor de F[1,4] de 100.9. La diferencia se debe a redondeo.

4. Supuestos de la prueba de t de Student.

Los mismos de ANDEVA: Muestreo Aleatorio, Independencia de Errores, Normalidad y Homogeneidad de Varianzas

5. Uso y abuso de la prueba de t-Student

La prueba de t de Student es una manera eficiente, rápida y poderosa de comparar la media de dos muestras para inferir diferencias entre poblaciones. En mi experiencia, el principal problema con la aplicación de la prueba de t-Student proviene del abuso de esta prueba para situaciones en las que se quiere comprar más de dos grupos.

Al parecer los investigadores que aprenden a usar solamente esta prueba piensan que ya que la aplican para comparar dos grupos pueden igualmente usarla para comparar, de a pares, tres o más grupos o niveles de un determinado factor. El problema es que la prueba esta diseñada para controlar la tasa de error Tipo I al realizar una comparación determinada. Si se quiere realizar todas las comparaciones posibles de un factor con cinco niveles, se tendrá $5! / 2! (5-2)! = 10$ comparaciones posibles. El problema es que la tasa de Error α es el error que puede cometerse por cada comparación, de manera que si se realizan cinco comparaciones el error total del estudio ya no es 5% si no que será mucho mayor.

Entonces, NO es posible aplicar la prueba de t-Student tantas veces como sea necesario para comparar varios (> 2) grupos.

III. PRUEBA DE T DE STUDENT PARA DIFERENCIAS ENTRE UNA MEDIA MUESTRAL Y UN PARAMETRO

En algunas circunstancias estamos interesados en comparar el valor de una muestra con un valor conocido de una población. Por ejemplo, imaginen que estamos estudiando una enzima determinada de la pared estomacal en una especie de pájaro introducida en el país y queremos ver si esta enzima se comporta igual que la enzima en el lugar de origen de los pájaros en Europa, en donde incidentalmente se han realizado muchos estudios con la enzima de esta especie. Entonces, decidimos comparar el tiempo transcurrido para alcanzar la mitad de la velocidad máxima de reacción o K_m (??) con los valores de esta constante observados en Europa, los que son igual 12.3. Para ello desarrollamos un experimento adecuado, en el cual medimos la velocidad de reacción de la enzima, NO en un macerado de la pared estomacal de un grupo de individuos, en cuyo caso perdemos la información de variación entre individuos y con ello toda replicación verdadera. Tampoco lo hacemos sobre varios macerados o muestras de un mismo individuo. Lo hacemos a través de tomar muestras independientes de los varios pájaros en Chile y comprar esta media muestral, Y , contra el valor poblacional de 12.3.

Nuestra hipótesis en este caso es:

$H_0: Y = \mu$ o, específicamente $H_0: Y = 12.3$

y la hipótesis alternativa sería $Y \neq 12.3$, por lo tanto nuevamente se trataría de un test de dos colas.

En este caso uno puede definir una prueba estadística como:

$$t_s = \frac{(\bar{Y} - m)}{\sqrt{(s^2 / n)}}$$

Luego comparamos este valor observado de t contra el valor esperado bajo la hipótesis nula con $n-1$ grados de libertad y un nivel de significancia preestablecido (0.05).

IV. METODOS DE DISTRIBUCION LIBRE PARA COMPARAR DOS MUESTRAS

Como les mencioné anteriormente, las pruebas de distribución libre son una alternativa cuando nuestros datos no cumplen con los supuestos de ANDEVA aún después de realizar transformaciones, a pesar de que en general los supuestos de ANDEVA son menos restrictivos de lo que algunos investigadores suponen. Cuando nuestro diseño es completamente aleatorio con un tratamiento y dos o más niveles, podemos recurrir a varias pruebas de distribución libre existentes. Sin embargo, la principal razón para usar estas pruebas de distribución libre debería ser cuando nuestros datos o hipótesis efectivamente estén expresadas en términos de rankings y deseamos realizar inferencias acerca de las medianas.

Al igual que la ANDEVA, algunas pruebas de distribución libre o no paramétricas someten a prueba hipótesis acerca de la *localización* de dos muestras, pero en este caso normalmente se usa la mediana como medida de localización. La mayoría de las pruebas de distribución libre para comparar muestras independientes (grupos) están basadas en la transformación de los datos originales a **rankings ordenados** (muchas veces traducido como “rangos”). Es decir, en vez de realizar nuestras comparaciones sobre las observaciones originales, primero debemos transformarlas a rangos desde el menor al mayor observado.

Bajo algunas circunstancias esto tiene algunas ventajas. Por ejemplo, si estamos interesados en comprar los tiempos de evacuación o emigración de dos especies de caracoles de áreas de parches sobre los que se ha removido el bosque, puede ser mucho más fácil registrar el orden en que individuos de una u otra especie desaparecen, que estar allí midiendo el tiempo exacto en que cada individuo dejó el parche. Por ejemplo, registramos como 1 el primer individuo de la especie 1 que dejó el parche, como 2 el segundo individuo de la especie 1, como 3 el tercer individuo de

la especie 2, etc. Estos datos expresados como rangos no pueden ser analizados con el método de análisis de varianza que vimos anteriormente.

Para analizar estos datos, existen varias alternativas:

Para comparar dos medias ($a = 2$):

- Prueba de las medianas ($H_0: \theta_1 = \theta_2$)
- Prueba U de Mann & Whitney
- Prueba de rangos sumados de Wilcoxon

Estas dos últimas pruebas son exactamente iguales y a veces se refiere a ellas como Wilcoxon-Mann-Whitney test.

Para comparar más de dos medias ($a > 2$, pero un sólo tratamiento):

- Prueba de las medianas para varios grupos
- Prueba de Kruscal-Wallis
- Prueba de Jonckheere-Terpstra

Todas estas pruebas se basan en la transformación de los datos a rankings, ordenando los datos del menor al mayor valor a través de todos los grupos a comparar y asignando el valor medio para resolver las igualdades o empates.

Las hipótesis están entonces no referidas a medias propiamente tal sino que a medianas u otros descriptores de localización.

Los procedimientos por lo general son muy fáciles de seguir y aplicar. Se calcula la sumatoria de los rankings y se compara con valores tabulados que entregan la probabilidad de obtener dichas sumatorias dado un determinado tamaño muestral.

Supuesto de estas pruebas de distribución libre:

- No requieren normalidad de los datos.
- Independencia de las observaciones
- Muestras aleatorias
- Homogeneidad de varianzas es necesaria
- Distribuciones de los grupos a comparar deben ser continuas y de la misma forma.

La prueba de Wilcoxon-Mann-Whitney aplicada al ejemplo de efectos de efluentes sobre tasas de crecimiento de peces en los acuarios se realizaría de la siguiente manera.

Primero que nada, debemos calcular los promedios de los seis peces en cada acuario antes de realizar la prueba de hipótesis pues la prueba de Wilcoxon – Mann-Whitney (y en general todas las pruebas de distribución libre) no nos permite incorporar jerarquía (submuestras) en nuestro diseño.

Luego, ordenamos todas las observaciones de menor a mayor y les asignamos un ranking del uno al 10. En este ejemplo y puesto que las tasas de crecimiento son una variable continua medida aquí con bastante precisión, no tenemos el problema de empates.

Tratam.	Acuario	Crecim.	Ranking	SUMA
1	1	104.93	10	
1	2	101.87	7	
1	3	103.19	9	
1	4	98.11	6	
1	5	103.04	8	$S_n = 40$
2	1	86.84	4	
2	2	84.99	2	
2	3	81.87	1	
2	4	88.71	5	
2	5	86.67	3	$S_m = 15$

Esto es todo lo que necesitamos calcular pues podemos ahora podemos comprar el valor menor de la sumatoria ($S_m = 15$) con el valor de una tabla en la cual nos dan los valores críticos para esta prueba y para 10 observaciones, a la significancia elegida ($\alpha = 0.05$) y para dos colas. Igualmente podemos usar el valor mayor de la sumatoria ($S_n = 40$). Es obvio que en teoría el valor de las sumatorias de rankings en cada grupo esperado debe estar alrededor de 27.5. Si la suma esta muy por arriba de este valor entonces un grupo tiene la mayoría de las observaciones más grandes y su mediana por lo tanto debe ser mayor. Si la suma esta muy por debajo, entonces la mediana de ese grupo debe ser menor que la del otro grupo.

Normalmente, no se comparan estos valores de sumatoria (S_m , S_n) directamente sino que se calcula la modificación de Mann-Whitney para lo cual existen más tablas disponibles. Con las sumatorias observadas y usando el tamaño muestral del grupo con menor sumatoria (m) o el de mayor sumatoria (n), se calcula U_m o U_n , respectivamente:

$$U_m = S_m - \frac{1}{2}m(m+1)$$

$$U_n = S_n - \frac{1}{2}n(n+1)$$

Luego se compara el valor de Um contra el valor de tabla.

Los resultados de este análisis para el ejemplo anterior, el cual si muestra normalidad y homogeneidad de varianza es el siguiente:

Wilcoxon Scores (Rank Sums) for Variable W
Classified by Variable A

A	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
1	5	40.0	27.5000000	4.78713554	8.0
2	5	15.0	27.5000000	4.78713554	3.0

Wilcoxon 2-Sample Test (Normal Approximation)
(with Continuity Correction of .5)

S= 40.0000 Z= 2.50672 Prob > |Z| = 0.0122

T-Test approx. Significance = 0.0335

Kruskal-Wallis Test (Chi-Square Approximation)
CHISQ= 6.8182 DF= 1 Prob > CHISQ= 0.0090

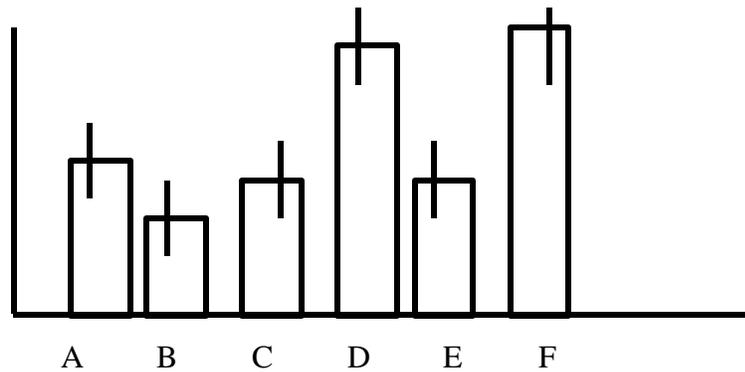
V. PRUEBAS PARA DETERMINAR QUE GRUPOS (MEDIAS) DIFIEREN SIGNIFICATIVAMENTE DE OTRAS

Luego de realizar un análisis de varianza y que hemos determinado que la hipótesis nula es falsa para un set varios niveles de una tratamiento ($a > 2$), enfrentamos un problema.

Si nuestro factor es aleatorio, entonces este es el fin del análisis pues hemos demostrado que la varianza debida a las diferencias entre niveles del tratamiento no es cero $\sigma^2 > 0$. Eso es todo lo que hemos especificado en la hipótesis: “diferencias en temperatura tienen efecto significativo sobre tasas metabólicas”.

En el caso de un factor fijo, sin embrago, el análisis de varianza usualmente no es el punto final. Por ejemplo, imaginen que estamos interesados en el efecto de las cinco drogas que se ofrecen en el mercado para reducir peso y luego de realizar un experimento en el cual seleccionamos 60 señoras gorditas, las pesamos y en forma completamente aleatoria les damos una de las 5 drogas a 10 señoras y al restante grupo de 10 les damos un placebo. Al cabo de tres semanas y de mantener a todas estas gorditas bajo exactamente la misma dieta de empanadas y costillar asado, las pesamos nuevamente y vemos la entonces las diferencias en peso bajo las distintas drogas y bajo el placebo.

Supongamos que los datos, expresados como pérdida de peso, se ven así:



Nuestro análisis de ANDEVA muestra que $P: >F_{0.05[5,54]} < 0.0001$

Ahora, es importante determinar cual de las drogas efectivamente difiere significativamente de las demás y si ellas al menos difieren del placebo. Para determinar esto, debemos realizar un análisis luego de realizar el ANDEVA.

Existen básicamente dos grandes tipos de pruebas que nos permiten saber que grupos o niveles de un tratamiento difieren significativamente de otros.

1.COMPARACIONES PLANEADAS O a priori

2.COMPARACIONES NO PLANEADAS O a posteriori

Clase 7: Pruebas *a priori* y *a posteriori*, Contrastes, Ortogonalidad

COMPARACIONES MÚLTIPLES PARA IDENTIFICAR LAS HIPÓTESIS ALTERNATIVAS

Como dijimos en la clase anterior, luego de realizar una ANDEVA y de encontrar diferencias significativas entre los niveles de un tratamiento, lo único que podemos decir es que al menos un grupo difiere significativamente, pero no sabemos cuál o cuales difieren de los otros.

Para determinar qué grupo o grupos (niveles de un factor) difieren significativamente, debemos realizar **comparaciones múltiples** luego de realizar el ANDEVA. Este tipo de comparaciones procede principalmente cuando el factor es fijo, debido a que se ha propuesto que algunos mecanismos explican cierto patrón en las observaciones. También se utilizan para determinar agrupaciones particulares de niveles de un factor aleatorio, pero muchas veces esto puede llevar a cuestionarse si nuestra clasificación del factor como aleatorio es correcta o deberíamos más bien interpretarlo como un factor fijo.

El realizar estas comparaciones múltiples equivale a especificar la hipótesis alternativa, es decir: determinar no solamente que al menos un grupo difiere, sino que ser específico respecto al patrón esperado.

Problema de excesivo error Tipo I

Identificar una hipótesis alternativa requiere varias comparaciones, pero si cada comparación es realizada con una probabilidad de error de Tipo I, α , el conjunto de comparaciones tiene un valor mucho más alto, i.e., se incrementa la probabilidad de rechazar H_0 , siendo ésta verdadera.

Por convención se ha establecido que α por experimento = 0.05, es la tasa de error total por experimento (= experimentwise error rate). Esto es, que la probabilidad que del total de comparaciones realizadas en un análisis, incluyendo el ANDEVA y todas las comparaciones múltiples, uno pueda cometer un error sea 1 en 20. Si en estudio o experimento determinado se aplican tres tratamientos o niveles de un factor, entonces existen tres comparaciones pareadas posibles y el investigador podría cometer 0, 1, 2 o 3 veces el mismo error, dependiendo de si todos los grupos a comparar han sido afectados por factores fuera de control del investigador o solamente algunos de los tres grupos. Por ejemplo, si el experimento consiste en examinar el efecto de tres hormonas sobre tasas de digestión de ratones, y los tres ratones inyectados con la hormona A tienden, por simple azar e independientemente de la hormona (factores no

controlados por el investigador) a digerir excepcionalmente muy rápido, los tres inyectados con la hormona B tienen a digerir excepcionalmente muy lento, mientras que los tres restantes bajo hormona C son ratones “promedio”. Al hacer la comparación entre A y B cometeremos un error, pues encontraremos diferencias que en la realidad no son producidas por las hormonas. Al comparar A y C nuevamente cometeremos un error por cuanto concluiremos que existen diferencias significativas cuando en la realidad no existen. Lo mismo al comparar B con C. Si se tiene un factor con siete niveles (ej. cinco enzimas, un control y un control de procedimiento), entonces tendremos 21 comparaciones pareadas posibles y de hecho, debemos esperar que por simple azar al menos una comparación sea significativa, aunque no existan diferencias entre ninguno de los tratamientos.

Entonces, al realizar comparaciones múltiples, lo que queremos controlar es que el **Experimentwise Error Rate** sea efectivamente de 1 en 20 (= 0.05): Que la probabilidad de cometer un error tipo I en todas las comparaciones realizadas con ese set de datos proveniente de un estudio o experimento determinado sea de 1 en 20.

Existen dos aproximaciones muy distintas para determinar que grupos difieren significativamente y esto depende de si las comparaciones a realizar son planeadas antes de realizar el experimento o son más bien sugeridas por los resultados después de realizar el experimento.

II. COMPARACIONES PLANEADAS O *a priori*

- Estas comparaciones son diseñadas y planeadas en forma *independiente de los resultados obtenidos*.
- Deben ser planeadas *antes* de realizar el experimento de acuerdo al interés específico de los investigadores.
- Las comparaciones a realizar NO pueden cambiarse después de realizado el experimento y de obtener los resultados.
- En general NO es posible comparar todos los grupos unos contra otros (existen restricciones).

III. EJEMPLOS DE COMPARACIONES PLANEADAS:

A. Un estudio para ver el efecto de dos drogas sobre el crecimiento de chanchitos incluye también un placebo o control. Entonces, los investigadores quieren saber:

- 1) Si existe un efecto significativo de las drogas (ANDEVA)

H₀: $\mu_1 = \mu_2 = \mu_3$ (1= control, 2= droga A, 3= droga B)

H_a: Al menos un μ diferente

2) Si las dos drogas difieren significativamente una de la otra

H₀ (primer contraste): $\mu_2 = \mu_3$

H_a: (primer contraste): $\mu_2 \neq \mu_3$

3) Si las dos drogas difieren significativamente del control

H₀ (segundo contraste): $\mu_1 = \mu_2 = \mu_3$

H_a (segundo contraste): $\mu_1 \neq \mu_2 = \mu_3$

B. Un estudio de herbívora por conejos en la zona semi-desértica de Chile central utiliza grandes jaulas para excluir los conejos de áreas cubiertas con vegetación. Los investigadores utilizan, además de las zonas control sin jaulas (con conejos), un “techo” del mismo material de la jaula, pero que permite el libre acceso de conejos. Entonces, los investigadores planean las siguientes comparaciones:

1) Si existe un efecto significativo del tratamiento (ANDEVA).

H₀: $\mu_1 = \mu_2 = \mu_3$ (1= control, 2= jaula exclusión, 3= techo)

H_a: Al menos un μ diferente

2) Si existe un efecto de la jaula misma (artefacto):

H₀ (c1, primer contraste): $\mu_1 = \mu_3$

H_a (c1, primer contraste): $\mu_1 \neq \mu_3$

3) Si existe un efecto de depredación por conejos:

H₀ (segundo contraste): $\mu_2 = \mu_3$

H_a: (segundo contraste): $\mu_2 \neq \mu_3$

IV. COMO REALIZAR COMPARACIONES PLANEADAS o *a priori*:

Para realizar comparaciones planeadas primero se construye un modelo e hipótesis biológicas específicas que identifican un patrón específico de diferencias entre tratamientos. Este patrón es definido o especificado como alternativo a la hipótesis nula y es determinado antes de realizar el experimento. Estas comparaciones planeadas o contrastes son comparaciones entre dos medias (dos niveles de un tratamiento), entre una media y un grupo de medias (ej. el promedio de varios grupos) o entre dos grupos de medias.

La secuencia u orden en la cual se realizaran las pruebas están especificados por las hipótesis definidas a priori. Esto garantiza un orden e interpretación más lógica que en pruebas a

posteriori. De hecho, algunas de las comparaciones planeadas podrán realizarse (para proseguir con la prueba del modelo biológico) solamente si no existen o existen diferencias en la comparación anterior.

Ejemplo:

Las diferencias entre sitios en el promedio de abundancia de hierbas es explicada por el modelo de herbivoría por conejos, que en algunas áreas causa una disminución en la abundancia observada de hierbas.

Nuestro modelo biológico entonces propone que la exclusión de depredadores debe conllevar a un cambio importante en la abundancia de hierbas. Para someter a prueba el modelo diseñamos un experimento de exclusión usando jaulas de malla que impiden el acceso a los conejos pero permiten el paso a organismos móviles de menor tamaño. Nuestra predicción es que la abundancia de hierbas será muy diferente dentro de las jaulas que en las áreas 'Control' en donde no se ha removido los conejos y se ha dejado la pradera intacta.

Puesto que el disponer jaulas en la pradera tiene un efecto importante de por sí, independiente de la exclusión de conejos (ej. sombra), debemos incluir además un "control de procedimiento". El control de procedimiento es lo que en experimentos de fisiología o medicina se denomina "placebo". El placebo o control de procedimiento es un tratamiento más que nos permite evaluar el efecto producido por artefactos de manipulación. En este tipo de estudios normalmente el principal efecto puede ser el proveer sombra, tanto para la vegetación como para otros organismos pequeños que pueden agregarse en las jaulas.

El experimento diseñado para evaluar nuestro modelo biológico consiste entonces de un Factor con tres tratamiento: Exclusión de conejos (E), Control sin ninguna manipulación (C) y un control de procedimiento en la forma de un Techo que provee sombra pero deja entrar a los conejos (CP).

A priori podemos plantear un resultado experimental interpretable:

- la abundancia media donde se ha excluido al depredador será mayor que en el control y en el control de procedimiento y
- las densidades en el control (C) y en el control de procedimiento (CP) son iguales
- Si efectivamente existen diferencias entre C y CP, entonces estamos usando una metodología que no permite obtener conclusiones sensibles acerca de la depredación. No podemos separar en forma inequívoca el efecto de depredadores y de sombra. Esto significa que primero debemos realizar esta última comparación: C versus CP

Así las hipótesis estadísticas son:

$$H_0: \mu_C = \mu_{CP} = \mu_E$$

$$H_a: \mu_C = \mu_{CP} \neq \mu_E$$

Donde C es control, CP es el control del procedimiento, y E la exclusión experimental
Otras posibilidades que no apoyan el modelo biológico:

$\mu_C < \mu_{CP} = \mu_E$ (implica un artefacto de la jaula sin efecto depredación)

$\mu_C < \mu_{CP} < \mu_E$ (implica artefacto y posible efecto depredación no discernible)

Secuencia para distinguir entre H_0 y la H_a específica:

1. Deben observarse diferencias entre medias en el ANDEVA, de otra forma H_0 se mantiene y no existe ninguna razón para realizar comparaciones múltiples, planeadas o no planeadas.
2. La media del control debe ser comparada con la media del control del procedimiento, y no deben encontrarse diferencias
3. Si la media del control no difiere de la media del control del procedimiento, entonces la media del control y control del procedimiento deben ser comparada con la media de la exclusión, y deben encontrarse diferencias en la dirección especificada,

Así, se ha especificado sólo dos comparaciones entre tratamientos y el orden a hacerlo a sido definido completamente *a priori*

Los contrastes (C) son una forma especial de funciones lineales que se definen a partir de los “coeficientes de contraste” (k_i) para cada grupo i a comparar y las medias de cada grupo μ_i

$$C = \sum_{i=1}^a k_i \mu_i = k_1 \mu_1 + k_2 \mu_2 + k_3 \mu_3 + \dots + k_a \mu_a$$

En donde sumatoria de todos los coeficientes de un contraste lineal (k_i) debe ser igual a cero: $C = \sum_{i=1}^a k_i = 0$. La combinación de coeficientes (k_i) en un contraste determinado se denomina “función lineal de contraste” (L). Una estimación de estos contrastes se obtiene al reemplazar las medias observadas (muestrales) por las medias poblacionales de cada grupo.

Los contrastes así definidos son una partición de la suma de cuadrados entre grupos correspondientes a ese factor en la ANDEVA principal.

Cada comparación realizada tiene 1 grado de libertad (por contraste lineal) como numerador y los mismos grados de libertad del Error de la ANDEVA como denominador. Esto significa que tenemos una manera muy “poderosa” de realizar comparaciones pues podemos usar todas las replicas del experimento.

Suponiendo que hemos encontrado diferencias significativas entre los grupos a comprara en nuestro ejemplo de herbivoría por conejos, necesitamos primero comparar el control contra el control de procedimiento. Para ello establecemos la siguiente función lineal

$$L_1: \begin{array}{ccc} \text{E} & \text{C} & \text{CP} \\ 0 & -1 & +1 \end{array} \text{ en donde } \sum k_i = 0$$

Esta función lineal con coeficientes $k_E = 0$, $k_C = -1$ y $k_{CP} = +1$ conforma el primer contraste al multiplicar por los promedios de los respectivos grupos:

$$C_1: 1 Y_C - 1 Y_{CP} + 0 Y_E$$

Esta es una comparación entre el primer y segundo nivel del factor “depredación” o tratamiento. Se puede usar cualquier valor de los coeficientes de contraste, mientras estos tengan el mismo valor y signos opuestos entre las medias o grupos de medias a comparar. La sumatoria del contraste es una estimación del “efecto” de esa comparación específica. Si no existen diferencias entre CP y C, entonces la ecuación de contraste debe ser cero.

Luego necesitamos realizar el segundo contraste, que estará definido por

$$C_2: 0.5 Y_C + 0.5 Y_{CP} - 1 Y_E \text{ o igualmente por:}$$

$$C_2: 1 Y_C + 1 Y_{CP} - 2 Y_E$$

En ambos casos la sumatoria de los coeficientes es igual a cero.

Esta es una comparación entre el promedio de los dos primeros niveles del factor y el último nivel (exclusión). Los factores con igual signo de los coeficientes de contraste son promediados y comparados contra el o los de signos opuestos.

V. CONTRASTES ORTOGONALES

Cierto tipo de contrastes tienen la propiedad de ser *ortogonales*. Esto es, la realización de un contraste no provee ninguna información respecto de otro u otros contrastes: son completamente independientes. Los contrastes ortogonales presentan varias ventajas, pues al ser independientes: 1) Cada comparación tiene igual poder que la ANDEVA principal y 2) La partición de la suma de cuadrados en el análisis de varianza y los contrastes son aditivas.

Los contrastes son ortogonales solamente si se cumple que:

- la suma de grados de libertad de todas las comparaciones que se desea realizar NO excede a los **a-1** grados de libertad *entre grupos*. Puesto que cada contraste tiene un

grado de libertad (siempre se comparan dos medias o dos grupos de medias), no se pueden hacer más de $a-1$ contrastes de un experimento determinado.

- las comparaciones planeadas son independientes: cada contraste debe someter a prueba una relación independiente entre las medias. Esto se cumple cuando a) la suma de los coeficientes de cada comparación es igual a cero y b) la suma de la multiplicación de todos los coeficientes de cada nivel es también cero.

Por ejemplo, si dos comparaciones planeadas que se desea realizar son:

$$C_1 : k_1Y_1 + k_2Y_2 + k_3Y_3 + k_aY_a$$

$$C_2: d_1Y_1 + d_2Y_2 + d_3Y_3 + d_aY_a$$

Entonces estos contrastes son ortogonales si:

$$\sum_{i=1}^a \frac{k_i d_i}{n_i} = 0$$

Si todos los grupos a comparar tienen igual tamaño muestral entonces la ortogonalidad se cumple

cuando $\sum_{i=1}^a k_i d_i = 0$

En el ejemplo de herbivoría por conejos y suponiendo que tenemos un diseño balanceado, nuestros contrastes son

	Exclusión	Control	Control Proced.	Suma
C_1	0	-1	+1	0
C_2	-2	+1	+1	0
$C_1 \times C_2$	0	-1	+1	0

Es decir nuestros contrastes son ortogonales y tendremos el máximo de poder para someter a prueba nuestro modelo. Si deseáramos realizar otro contraste, por ejemplo comparar el Control contra la Exclusión separadamente, este contraste sobrepasaría el número permitido ($a-1 = 2$) y no sería ortogonal.

Si en lugar de estos dos contrastes nuestro modelo (otro distinto al planteado) sugiere comprar primero el control contra el control de procedimiento y luego el control contra la Exclusión, tendremos solamente dos comparaciones, pero los contrastes no son ortogonales. Demuestra que no son ortogonales.

IV. PARTICION DE SUMA DE CUADRADOS Y PRUEBA DE HIPOTESIS PARA CONTRASTES

Es posible calcular la suma de cuadrados para cada contraste que desea realizar. Es suma de cuadrados de contrastes (SCC_i) se calcula como:

$$SSC_i = \frac{\left(\sum_{i=1}^a k_i \bar{y}_i \right)^2}{\sum \left(\frac{k_i^2}{n_i} \right)}$$

Si el diseño es balanceado y todos los n_i son iguales, entonces la SCC se reduce a:

$$SSC_i = \frac{n \left(\sum_{i=1}^a k_i \bar{y}_i \right)^2}{\sum (k_i^2)}$$

Puesto que cada contraste tiene un grado de libertad los cuadrados medios de cada contraste son: $CMC_1 = SCC_1/1 = SCC_1$

La hipótesis para todos los contrastes son de la forma:

$$H_0: C_1 = 0\mu_1 + \mu_2 - \mu_3 = 0$$

$$H_0: C_2 = 2\mu_1 - \mu_2 - \mu_3 = 0$$

Las hipótesis alternativas son de la forma $H_a: C_1 \neq 0$

Las Pruebas de Hipótesis de todos los contrastes se realizan a través de dividir el cuadrado medio de cada contraste por el cuadrado medio del error experimental del ANDEVA:

$$F_{C_1} = CMC_1 / CME$$

Esta prueba de F tendrá un grado de libertad en el numerador y los mismo grados de libertad del error experimental en el denominador.

Es obvio que puestos que en cada contraste se comparan dos grupos (dos medias o dos grupos de medias), la prueba de hipótesis de cada contraste puede hacerse con una prueba de t-Student.

V. EJEMPLO DE CONTRASTE ORTOGONALES:

1. Experimento de herbivoría: Jaulas de Exclusión, Techo y Control

La Tabla 7.1 presenta los resultados del experimento de herbivoría, el se realizó con cuatro réplicas para cada tratamiento:

Tabla 7.1. TRT =1 es la Exclusión, TRT = 2 es el Techo y TRT = 3 es el Control

	TRT	REP	PLANTAS
	1	1	130. 122
	1	2	144. 528
	1	3	148. 312
	1	4	144. 065
	2	1	131. 582
	2	2	119. 600
	2	3	122. 564
	2	4	119. 346
	3	1	112. 182
	3	2	128. 940
	3	3	119. 705
	3	4	119. 875

PROMEDIOS POR TRATAMIENTO:			
TRT	MEDIA	EE	n
1	141. 757	3. 99308	4
2	123. 273	2. 86444	4
3	120. 176	3. 42804	4

En la Tabla 7.2 Se presentan los resultados del análisis completo de los resultados de 1 experimento de Herbivoría, incluyendo el ANDEVA y los contrastes planeados. Primero se evaluó el contraste entre el Control y el Control de Procedimiento (Techo) y luego el contraste entre la Exclusión y el promedio de los otros dos. Estos son los resultados del programa SAS.

Cada contraste tiene un grado de libertad y los dos suman al total de contrastes posibles. Puesto que los dos contrastes seleccionados son ortogonales, la suma de las sumas cuadrados de los dos contrastes es igual a la Suma de Cuadrados de Tratamiento (entre grupos) del ANDEVA.

Tabla 7.2 Output de SAS para el experimento de Herbivoría.

Dependent Variable: PLANTS	Sum of	Mean
----------------------------	--------	------

Source	DF	Squares	Square	F Value	Pr > F
Model (TRT)	2	1089.31808	544.65904	11.38	0.0034
Error	9	430.81453	47.86828		
Corrected Total	11	1520.13262			
	R-Square	C. V.	Root MSE	PLANTS Mean	
	0.716594	5.388311	6.91869	128.402	
Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
Techo vs. Control	1	19.19075	19.19075	0.40	0.5424
Excl vs. Otros	1	1070.12733	1070.12733	22.36	0.0011

Un buen ejercicio es calcular a mano las sumas de cuadrado de cada contraste usando los valores de la media de la Tabla 7.1. Hacer este contraste con una calculadora como ejercicio y buscar el valor de significancia de F en una tabla con valores tabulados.

VI. CURVAS DE RESPUESTA

Muchas veces la variable independiente o tratamiento es una variable continua cuantitativa, como por ejemplo la concentración de oxígeno extra celular, o la concentración de enzima catalizadora al interior de la célula. En este tipo de estudios resulta de interés ver la relación cuantitativa entre la variable respuesta, como puede ser la producción de ATP o la producción de proteína, y la variable independiente. En estos casos resulta útil estudiar las curvas de respuesta (o superficie respuesta en el caso de diseños factoriales), a través de ajustar coeficientes de contraste polinomiales.

VII. CONTRASTES NO ORTOGONALES

Como mencionamos anteriormente, los contrastes a realizar deben ser dictados por las preguntas del investigador y no por las propiedades estadísticas de ellos. A veces los investigadores desea realizar contrastes que no son ortogonales y en estos casos se debe corregir la tasa de error por comparación α_c de tal manera que la tasa de error por experimento α_e sea efectivamente el valor deseado de probabilidad (0.05). El método de Dunn-Sidák es la prueba más poderosa que controla la tasa de error por experimento para un número de comparaciones o contrastes planeados no ortogonales. Otra posibilidad, menos poderosa pero más simple es usar Pruebas de t de Bonferroni (con corrección de Bonferroni).

VI. COMPARACIONES NO PLANEADAS O *a posteriori*

- Al contrario de las comparaciones planeadas, las comparaciones *a posteriori* son realizadas después de realizado el experimento y de que se conocen los resultados.
- Estas comparaciones son sugeridas por los resultados mismos y NO son planeadas antes del experimento por los investigadores.
- Puesto que las comparaciones se realizan después de conocer los resultados, las pruebas de hipótesis deben considerar el hecho que ya no se trata de una muestra completamente aleatoria de una población normal, sino que de una muestra selectiva o parcial.

V. EJEMPLOS DE PRUEBAS *a posteriori*

Un estudio investiga el efecto de seis diferentes químicos sobre la velocidad de natación de insectos. No existen razones antes de realizar el estudio para querer comparar un químico en particular contra otro. Luego de realizado el estudio y que los investigadores encuentran diferencias significativas, ahora quieren saber que químico o grupos de químicos tienen efecto significativo y cuales no. Las hipótesis que someten a prueba son:

- 1) Si existe un efecto significativo de las drogas (ANDEVA)

$$\mathbf{H_0:} \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6$$

$\mathbf{H_a:}$ Al menos un μ diferente

- 2) La hipótesis particular dependerá de los resultados. Normalmente se realizan “todas las comparaciones interesantes” entre medias y/o entre grupos de medias.

VI. PRUEBAS *a posteriori*

Existe un gran número de pruebas estadísticas *a posteriori* que utilizan diferentes métodos y filosofías para calcular niveles de significancia. En general los libros de estadística entregan pocos antecedentes acerca de estos métodos y los investigadores los usan en forma más bien antojadiza. En muchas circunstancias se usan estas pruebas de comparaciones no planeadas cuando se podrían haber usado pruebas de comparaciones planeadas.

La mayoría de las pruebas *a posteriori* (no todas) tratan de mantener constante lo que se llama “*tasa experimental de error tipo I*” (experimentwise error rate, $\mathbf{\alpha_e}$): Mantener constante el nivel de significancia *alfa* (de cometer un error Tipo I) en la ANDEVA y en todas las pruebas y comparaciones entre medias o grupos de medias.

Para mantener constante la Tasa de Error por Experimento (α_e) se debe reducir la tasa de error por comparación (α_c). La probabilidad de cometer un número determinado de errores (x)

en b comparaciones, es decir la Tasa de Error pro Experimento, si las comparaciones son independientes (sin usar los mismos datos más de una vez) es:

$$P(x) = \frac{b!}{x!(b-x)!} \mathbf{a}_c (1 - \mathbf{a}_c)^{b-x}$$

Así, la probabilidad de cometer un error ($x = 1$) en tres comparaciones a realizar ($b = 3$) es:

$$P(x = 1) = \mathbf{a}_c = 1 - (1 - \mathbf{a}_c)^b$$

Si hemos mantenido un α_c de 0.05 por cada comparación que realizamos en un experimento (ej. aplicando un test de t para comparar tres niveles de un factor de a pares), entonces la tasa real de Error Tipo I del experimento es de: $\alpha_e = 0.143$

Si mantenemos constante la tasa de error por experimento a nivel de por ejemplo el 5% ($\alpha_e = 0.05$), entonces resulta obvio que debemos disminuir la tasa de error por comparación. La relación es la siguiente:

$$\mathbf{a}_c = 1 - (1 - \alpha_e)^{1/b}$$

Si hemos decidido realizar el experimento con una tasa de error por experimento de 0.05 y realizaremos tres comparaciones, entonces cada comparación tendrá una tasa de error de 0.017.

Si se usan los mismos datos para realizar varias comparaciones, entonces la tasa de error será un poco más baja, debido a la correlación entre los datos. Los valores calculados con la ecuaciones de arriba se pueden considerar como el límite superior de la tasa de error.

Resulta obvio que mientras más comparaciones se realicen, menor es el α_c o nivel de significancia de cada comparación en particular y por ello las pruebas *a posteriori* son conservadoras. Todas las pruebas *a posteriori* tienen menor poder que una prueba ANDEVA y menor poder que los contrastes planeados ortogonales. De hecho, uno de los principales problemas de las pruebas a posteriori es aumentar el poder de la prueba.

- La gran mayoría de los métodos se basa en el calculo del valor crítico de *Rangos Estudentizados* (Studentized Ranges), \mathbf{Q}_a . Otros usan valores de F o t-Student.

El tópico de comparaciones múltiples no planeadas excede por mucho lo que podemos hacer en esta clase. La mejor revisión del tópico que yo conozco es la de

Day, R.W. & Quinn, G. P. (1989). *Ecological Monographs* 59:433-463.

Uno de los problemas de la aplicación de estas pruebas múltiples no planeadas es que pone demasiado énfasis en la significancia de los tratamientos y se aleja de la o las preguntas biológicas de interés así como de el tamaño de los efectos

- Algunos métodos son para comparación entre medias pareadas:

Prueba LSD (mínima diferencia significativa)

Prueba de Dunn-Sidak

Prueba T

Prueba T'

Prueba GT2

Prueba de Tukey (HSD)

Prueba de Welsh

Prueba Student-Newman-Keuls (SNK)

- Algunos métodos para todas las comparaciones posibles:

Prueba de Scheffé

Prueba T

Prueba GT2

SS-STP

Prueba LSD

La prueba LSD (mínima diferencia significativa) trata las comparaciones individuales como la unidad de interés y por ello usa la Tasa de Error por comparación. Debe usarse solamente cuando el F de la ANDEVA muestra diferencias significativas. Entonces se dice que la prueba esta "protegida". NO debe usarse si no se encuentran diferencias en la ANDEVA.

El método se basa en calcular para cualquier par de medias observadas la cantidad:

$$LSD(\alpha) = t_{\alpha/2, v} \sqrt{s^2 \left[\frac{1}{n_i} + \frac{1}{n_j} \right]}$$

La hipótesis nula, $H_0: \mu_i = \mu_j$ se rechaza si $|y_i - y_j| > LSD$

La prueba HSD de Tukey

La prueba HSD (Diferencia Honestamente Significativa) de Tukey trata al set de comparaciones completa, es decir el experimento mismo, como la unidad de interés y por ello trata de controlar la Tasa de Error Experimental. La prueba se basa en el cálculo de los **rango estudentizado, q**:

$$q = \frac{\bar{y}(\text{mayor}) - \bar{y}(\text{menor})}{\sqrt{\frac{s^2}{n}}}$$

Con este valor se calcula el valor de HSD:

$$HSD(a, \mathbf{a}_e) = q_{a, a, v} \sqrt{\frac{s^2}{n}}$$

La hipótesis nula, $H_0: \mu_i = \mu_j$ se rechaza si $|y_i - y_j| > HSD(a, \alpha_e)$

Clase 8: ANDEVA ANIDADOS

I. DISEÑOS ANIDADOS O CON JERARQUÍA

El diseño CRD con submuestras es un diseño anidado en el cual se reconoce la existencia de varios niveles dentro de cada nivel de las unidades experimentales.

En general, un diseño anidado es cuando se utiliza replicación de unidades experimentales en al menos dos niveles de una jerarquía. Cada nivel de un tratamiento, fijo o aleatorio, tiene representación en cada uno de los niveles del otro tratamiento

Modelo lineal para ANDEVA anidado:

$$Y_{kij} = \mathbf{m} + \mathbf{T}_k + \mathbf{J}_{ki} + \mathbf{e}_{kij}$$

También se puede representar como: $Y_{kij} = \mathbf{m} + \mathbf{T}_k + \mathbf{J}_{i(k)} + \mathbf{e}_{j(i(k))}$

Ejemplo 1:

Experimento para ver la concentración de metales pesados en plantas de tomate provenientes de los 5 productores de tomate ($k = 5$ parcelas) en la zona de Ventanas, de cada parcela se seleccionan 5 plantas al azar ($i = 5$ plantas de tomate por parcela) y de cada planta se recolectan 5 tomates seleccionados al azar ($j = 5$ tomates por planta).

En estos casos es mucho más general y más claro hablar de “factores anidados”, en vez de la terminología de “replicas” y submuestras pues cada nivel de la jerarquía de fuentes de variación es una fuente de replicación para el nivel superior.

Con un diseño como el descrito podemos responder hipótesis relacionadas a 1) la variación en concentraciones de metales pesados entre parcelas, y 2) variación entre plantas dentro de cada parcela. En este diseño no podemos someter a prueba hipótesis acerca de la variación entre tomates dentro de plantas pues NO tenemos replicación a este nivel. Es decir, no existe una fuente de variación aleatoria *dentro* del nivel tomates.

Sin embargo, podemos realizar un estudio de la contribución relativa a la varianza total en concentración de metales pesados a través de estudiar los componentes de varianza. En este caso si podemos estimar el porcentaje de contribución a la varianza total de los tres niveles (parcela, planta y tomates) en la jerarquía.

Ejemplo 2:

Estamos interesados en estudiar el grado de variabilidad en la composición de DNA satelital de mitocondrias de una especie de pez de río. Para poder interpretar estos patrones de variación en términos de radiación, diseñamos un muestreo jerárquico que incluye: 8 Ríos Principales seccionados al azar desde un mapa con el rango de distribución de la especie. En cada río existen varios contribuyentes secundarios y decidimos seleccionar 5 de estos en cada río, luego debemos seleccionar los esteros de cada afluente principal y decidimos muestreo 4 de estos esteros. Dentro de cada estero recolectamos 10 peces y de cada individuo obtenemos 4 muestras.

Nuestro diseño es:

ANDEVA

Fuente de Variación	g.l.
Entre Ríos	$a-1 = 8-1 = 7$
Entre Afluentes(Ríos)	$a(b-1) = 8(5-1) = 32$
Entre esteros(afluentes, ríos)	$ab(c-1) = 8*5(4-1) = 120$
Entre peces(esteros, afluentes, ríos)	$abc(d-1) = 8*5*4(10-1) = 1440$
entre muestras(peces, esteros, afluentes, ríos)	$abcd(e-1) = 8*5*4*10*(4-1) = 4800$

¿Cuál es la fuente de Variación apropiada para someter a prueba a hipótesis de variación significativa en composición (o variabilidad) de ADN entre Ríos (los más separados geográficamente)?

Es aquella fuente de variación inmediatamente inferior, la cual representa estimaciones *independientes* de la variación dentro de cada río. Esto es siempre cierto cuando los factores anidados son aleatorios.

La determinación de cuál es la fuente de variación apropiada para someter a prueba una determinada hipótesis se basa en la forma Cuadrados Medios Esperados. En los diseños anidados:

1. El último termino en el modelo, es decir la fuente de variación “*residual*” o *Error* experimental debido a las unidades experimentales más pequeñas debe ser un factor aleatorio. Esto es valido para cualquier diseño experimental.

Esta fuente de variación mide (debe medir) la variación entre unidades experimentales similares pero independientes (replicas o submuestras dentro de otras unidades experimentales). La variación esta dada por el *azar*, es decir la varianza no explicada en el modelo.

Esta fuente siempre serán réplicas para algún factor superior en el modelo jerárquico.

2. El “Tratamiento” o Primer factor en el diseño anidado (primero en la jerarquía) puede ser un factor fijo o aleatorio.

Por ejemplo: Los cinco productores de tomates en ventanas es un factor fijo (si no hay más productores). Los ocho ríos para el estudio de DNA de peces es un factor aleatorio (ríos seleccionados al azar).

3. Los factores anidados en teoría pueden ser fijos o aleatorios. Sin embargo, en general los factores son aleatorios (existen muy pocos ejemplos de factores fijos) y para todos los efectos prácticos podemos decir que casi siempre son aleatorios.

Modelo para ejemplo de peces:

$$Y_{kijhm} = \mu + A_k + B_{ki} + C_{kij} + D_{kijh} + e_{kijhm}$$

Cuadrados medios esperados en ANDEVA anidado, factores aleatorios:

Fuente de Variación	CME
Entre Ríos	$\sigma_e^2 + e \sigma_D^2 + ed \sigma_C^2 + edc \sigma_B^2 + edcb \sigma_A^2$
Entre Afluentes(Ríos)	$\sigma_e^2 + e \sigma_D^2 + ed \sigma_C^2 + edc \sigma_B^2$
Entre esteros(afluentes, ríos)	$\sigma_e^2 + e \sigma_D^2 + ed \sigma_C^2$
Entre peces(esteros, afluentes, ríos)	$\sigma_e^2 + e \sigma_D^2$
Entre muestras(peces, est., aflue., ríos)	σ_e^2

Una de las características más interesantes de estos diseños es que uno puede someter a prueba la hipótesis de significancia de cualquier nivel, *independientemente* de si existen diferencias significativas o no entre los niveles de la fuente de variación inferior. Por ejemplo, se puede someter a prueba la hipótesis de variaciones significativas entre afluentes aunque se haya encontrado que los esteros dentro de afluentes difieren significativamente.

Si el diseño no es balanceado (igual número de replicación dentro de cada nivel), no existen problemas con calcular un valor de F, pero la probabilidad asociada puede dar valores aproximados. Es ideal tener el mismo nivel de replicación dentro de cada nivel en la jerarquía.

Entonces, basados en los CME de la tabla de arriba, debemos realizar los siguientes cálculos de F para someter a prueba las distintas hipótesis:

Fuente de Variación	CM	F
Entre Ríos	CMríos	CMríos / CMAflu.
Entre Afluentes(Ríos)	CMAflu.	CMAflu./ CMesteros
Entre esteros(afluentes, ríos)	CMesteros	CMesteros/ CMpeces
Entre peces(esteros, afluentes, ríos)	CMpeces	CMpeces / CMresid.
Entre muestras(peces, est., aflue., ríos)	CMmuestras CMresidual	=

Estimación de Componentes de Varianza

Una de las principales razones para realizar diseños anidados es la estimación de la contribución de los distintos factores a la varianza total observada en la varianza respuesta. En la mayoría de los casos estaremos interesados en determinar cual es la contribución específica a la varianza debida a un factor particular en la jerarquía (ej. $\sigma_{\text{ríos}}^2$, $\sigma_{\text{esteros}}^2$).

La estimación puntual de estos componentes de varianza se realiza a través de igualar los Cuadrados Medios con sus Cuadrados Medios Esperados, por ejemplo:

$$\text{CM ríos} = \sigma_e^2 + e \sigma_D^2 + ed \sigma_C^2 + edc \sigma_B^2 + edcb \sigma_A^2$$

$$\text{CM afluentes} = \sigma_e^2 + e \sigma_D^2 + ed \sigma_C^2 + edc \sigma_B^2$$

$$\text{CM esteros} = \sigma_e^2 + e \sigma_D^2 + ed \sigma_C^2$$

$$\text{CM peces} = \sigma_e^2 + e \sigma_D^2$$

$$\text{CM muestras (resid)} = \sigma_e^2$$

Si deseamos estimar la varianza debida a ríos ($\sigma_{\text{ríos}}^2$), entonces debemos realizar las subtracciones apropiadas:

$$\sigma_{\text{ríos}}^2 = \frac{\text{CM ríos} - \text{CM afluentes}}{\text{edcb}}$$

Lo mismo para los otros componentes de varianza:

$$\sigma_{\text{afluentes}}^2 = \frac{\text{CM afluentes} - \text{CM esteros}}{\text{edc}}$$

$$\sigma_{\text{esteros}}^2 = \frac{\text{CM esteros} - \text{CM peces}}{\text{ed}}$$

$$\sigma^2_{\text{peces}} = \frac{\text{CM peces} - \text{CM residual}}{e}$$

$$\sigma^2_{\text{muestras}} = \text{CM residual}$$

$$\mathbf{S}^2_{\text{TOTAL}} = \mathbf{S}^2_{\text{ríos}} + \mathbf{S}^2_{\text{afuentes}} + \mathbf{S}^2_{\text{esteros}} + \mathbf{S}^2_{\text{peces}} + \mathbf{S}^2_{\text{muestras}}$$

El cálculo de componentes de varianza se facilita también en un diseño balanceado. Aquí es muy importante determinar si el factor principal es fijo o aleatorio pues el ajuste de términos puede hacerse con o sin incorporar este factor en primer lugar en el modelo.

Existen varias otras técnicas para estimar componentes de varianza que no están basadas en la técnica de cuadrados mínimos usada por ANDEVA y explicada arriba. El problema de la técnica de cuadrados mínimos es que debido a la formulación matemática, es posible encontrar valores negativos de varianza luego de despejar las ecuaciones. Obviamente, las varianzas NO existen, de manera que esto es solamente un error de cálculo. Bajo estas circunstancias algunos autores sugieren simplemente hacer todas las varianzas negativas iguales a cero. Esto no tiene mucho efecto en los resultados por cuanto si las varianzas resultan negativas usando los cuadrados mínimos, entonces estas deben ser en la realidad muy cercanas a cero.

Otra manera de estimar los componentes de varianza es a través de procesos de iteración y de máxima verosimilitud. Estas técnicas básicamente intentan encontrar el mejor ajuste al modelo lineal de componentes de varianza.

Clase 9: Diseños Factoriales

I. RECAPITULACIÓN.

Hasta este momento hemos visto dos tipos de diseños experimentales:

Diseño Completamente Aleatorio.

- El más simple de todos los diseños.
- Para el caso de dos grupos a comparar se puede usar prueba t-Student
- Varias “soluciones” no paramétricas
- En ANDEVA se pueden comprar varios niveles de un mismo tratamiento
- La prueba de hipótesis es una sola (un solo F) y no varía si el factor es fijo o aleatorio

Diseño Anidado o Jerárquico

- En el caso más simple es un diseño CRD con submuestras
- En cada nivel de un factor, existen varios niveles de otro factor (anidado)
- Pueden haber varios niveles en la jerarquía
- El primer factor principal (más alto en la jerarquía) puede ser fijo o aleatorio
- Los factores anidados son casi siempre aleatorios
- Existen varias pruebas de hipótesis (varios F): Uno para cada fuente de variación
- Cada fuente de variación anidada sirve como “error” para el factor superior
- Pruebas de hipótesis a cada nivel son independientes de significancia a nivel inferior

I. ANALISIS DE VARIANZA DE DOS VIAS

En los análisis CRD tenemos solamente un tratamiento o factor el cual tiene dos o más niveles o grupos. Todos estos niveles son cualitativamente iguales:

Ejemplos de factores en ANDEVA *simple o de una vía*

tratamiento/factor	Niveles/Grupos
“hormona crecimiento”	con o sin droga
“drogas”	tres drogas distintas
“sitio”	seis sitios elegidos al azar
“temperatura”	cinco valores de temperatura
“sexo”	machos y hembras

A veces estamos interesados en el efecto simultáneo de dos factores o tratamientos sobre una variable determinada. Cada uno de estos tratamientos o factores puede tener dos o más niveles y NO están anidados un dentro del otro.

“Ambiente”	“Genotipo”	
	Tipo-A	Tipo-B
14°	Tipo-A, 14°	Tipo-B, 14°
23°	Tipo-A, 23	Tipo-B, 23
30°	Tipo-A, 30	Tipo-B, 30
32°	Tipo-A, 32	Tipo-B, 32

“Altura Marea”	“Depredador”	
	Con	Sin
BAJO	Con, Bajo	Sin, Bajo
MEDIO	Con, Medio	Sin, Medio
ALTO	Con, Alto	Sin, Alto

En estos casos el análisis apropiado es un ANDEVA de *Dos Vías*, el cual considera los dos factores o tratamientos al mismo tiempo.

La única diferencia importante entre el ANDEVA de una y el de dos vías es que ahora tenemos una fuente de variación “*extra*”, NO existente en los Análisis de varianza separados de una vía: esta fuente de variación se llama **INTERACCIÓN** entre los tratamientos.

Ejemplo:

- Se quiere ver el efecto de estrellas de mar (depredador) sobre la densidad de choritos cuando estos se encuentran cubiertos o no por un dosel de algas (escondidos por las algas o no).

Tratamiento A (Principal): Presencia de estrellas depredadores.

Niveles: Con o Sin depredador ($a = 2$)

Hipótesis: Ho1: No hay efecto de las estrellas sobre los choritos

Ha: Si existe efecto de las estrellas sobre los choritos

Tratamiento B (Principal): Presencia de dosel de algas cubriendo los choritos

Niveles: Con o Sin dosel de algas ($b = 2$)

Hipótesis: Ho2: No hay efecto (directo) del dosel de algas sobre choritos

Ha: Si existe un efecto (directo) del dosel de algas sobre los choritos

Interacción: El efecto conjunto de las estrellas y dosel de algas sobre los choritos

Niveles: Se analizan las diferencias entre niveles de un tratamiento con respecto a los niveles del otro tratamiento

Hipótesis: Ho3: El efecto de las estrellas es independiente de la presencia de un dosel e algas

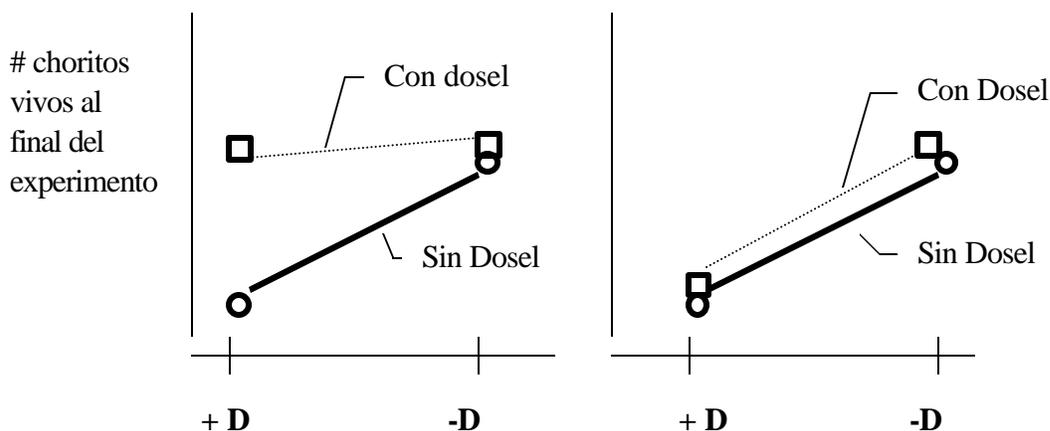
Ha: El efecto de las estrellas depende de la presencia de dosel algas.

La existencia de Interacción Significativa Significa que el efecto de un tratamiento sobre la variable en cuestión varía dependiendo de los niveles del otro tratamiento. Los efectos NO son independientes.

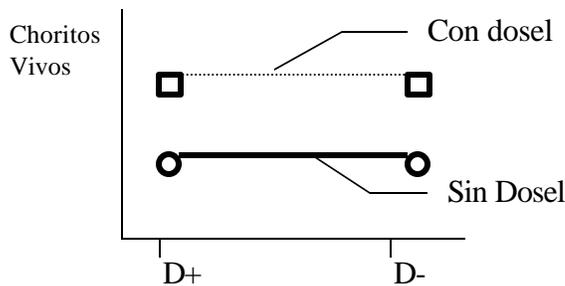
La interacción significativa significa que un tratamiento depende del otro tratamiento. Puesto que la evaluación de las hipótesis Ho1 y Ho2, acerca de los “efectos Principales”, supone que estos tratamientos son independientes, el supuesto de la no interacción debe ser evaluado antes de verificar las hipótesis acerca de los tratamientos.

Si la interacción NO es significativa, entonces se puede explorar cuál es el efecto de los tratamientos (Ho1 y Ho2).

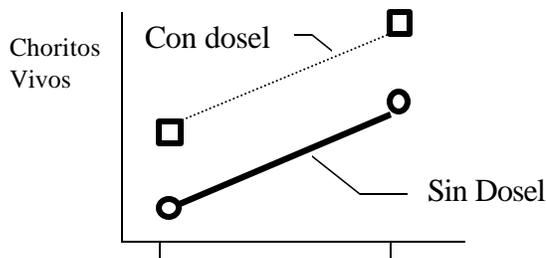
Si la interacción es significativa, entonces no tiene sentido evaluar las hipótesis acerca de los efectos principales, ya que se ha demostrado que estos tratamientos no son independientes.



A. Efecto significativo de dosel pero No efecto de depredadores Ni Interacción



B. Efecto de depredadores y efecto de dosel, pero no interacción



C. No efecto de depredadores ni efecto de dosel, pero efecto de interacción (tarea).

Muchas veces, es precisamente la interacción entre tratamientos la que tiene mayor significado biológico.

- Por ejemplo, cuando estamos interesados en la expresión fenotípica de diferentes genotipos, estamos realmente interesados en como el genotipo interactúa con variables ambientales. En estos casos el saber si los tratamientos principales son o no significativos no es de mucho interés.
- Cuando queremos saber el efecto de dos hormonas administradas a diferentes horas del día (mañana, tarde y noche), en realidad estamos interesados en saber si una “combinación” específica de hormonas y hora del día tiene mayor efecto. La hormona A es mejor que la hormona B “ cuando se administra por la tarde” .
- Si queremos saber el efecto de la instalación de una planta nuclear sobre las población de locos en la costa de Chile. Tenemos bahías similares, una de las cuales es designada como control y la otra es donde se instalara la planta. tenemos datos de antes que se instalara planta y después que se instalara la planta. Entonces, ¿Qué

fuente de variación nos muestra que la planta tiene un efecto significativo sobre la abundancia de locos?

El “tratamiento” antes/después => Sólo nos muestra posibles cambios temporales.

El tratamiento Bahía 1(control)/Bahía 2 (planta) => Sólo nos muestra posibles diferencias entre las bahías.

Solamente la interacción entre los tratamientos nos muestra el efecto de la planta.

Modelo Lineal para ANDEVA de Dos Vías Factorial:

$$Y_{kij} = \mu + A_k + B_i + AB_{ki} + e_{kij}$$

donde μ es la media de todas las poblaciones muestreadas,

Y_{kij} es la observación de la j replica, en el nivel k del factor A y en el nivel I del factor B

A_k representa el efecto del nivel k del factor A sobre la media

B_i representa el efecto del nivel I del factor B sobre la media

AB_{ki} representa la interacción para la combinación del nivel k de A y nivel I de B

e_{kij} es el error residual

Para el caso de ANDEVA de dos vías **Modelo I (A y B FIJOS)**, la Tabla de ANDEVA se ve así.

Fuente de Variación	g.l.	SC	CM	CME	F
Tratamiento A (T1)	a-1	SC _{T1}	SC _{T1} /a-1	$\sigma_e + bnT_A^2$	$\frac{CM_{T1}}{CM_{Err}}$
Tratamiento B (T2)	b-1	SC _{T2}	SC _{T2} /b-1	$\sigma_e + anT_B^2$	$\frac{CM_{T2}}{CM_{Err}}$
Interacción (A*B)	(a-1)(b-1)	SC _{Int}	SC _{Int} /(a-1)(b-1)	$\sigma_e + nT_{AB}^2$	$\frac{CM_{Int}}{CM_{Err}}$
Error (dentro de grupos)	ab (n-1)	SC _{Err}	SC _{Err} /ab(n-1)	σ_e	

La única manera de estimar la existencia de interacción es a través de un diseño *Ortogonal y Replicado*.

Un diseño *ortogonal* es aquel en el cual todos los niveles de un factor se combinan con todos los niveles del otro factor.

Cuando un diseño factorial no tiene réplicas verdaderas (réplicas de cada combinación de los niveles de los tratamientos), entonces no se puede determinar si la interacción entre tratamientos es significativa o no. Se debe entonces *asumir* que no existe interacción para poder someter a prueba las hipótesis de los efectos principales.

Es necesario someter a prueba la hipótesis Ho3 de la interacción *antes* de someter a prueba las hipótesis acerca de los efectos separados de los factores A y B. Si existe interacción entre los tratamientos esto significa que son dependientes y no es lógico proceder con las pruebas de hipótesis Ho1 y Ho2.

II. ¿POR QUÉ DEBEMOS USAR DISEÑOS FACTORIALES?

Existen dos razones importantes para usar diseños factoriales por sobre diseños de una vía separados. La primera es la información contenida en la interacción entre tratamientos. La segunda es la mayor eficiencia y poder de los diseños factoriales por sobre diseños de una vía separados.

1. Información acerca de las interacciones:

La mejor manera de ver la importancia de la información contenida en las interacciones es ver unos ejemplos.

Ejemplos:

1. La abundancia de la planta *Quercus chilensis* se relaciona negativamente con la abundancia local de la planta *Litreaea cáustica*, en todos los sitios en que se han observado estas especies en Chile central. Como explicación a este patrón de distribución, planteamos que existe competencia inter-específica por recursos (ej. nutrientes) entre estas especies y ello da cuenta de la correlación negativa en todos los sitios observados.

Este modelo explicativo requiere ser sometido a prueba por cuanto existen otras explicaciones que pueden dar cuenta de este patrón (ej. diferentes requerimientos de hábitat, competencia aparente). La predicción de este modelo es que la remoción experimental de la especie *Litraea* producirá un aumento en la abundancia de *Quercus*.

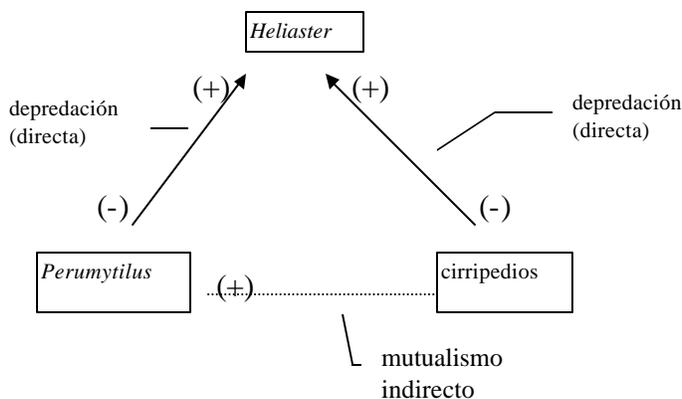
El experimento podría realizarse en un solo sitio, pero la observación original es que esta correlación ocurre en varios sitios, sin importar las diferencias en otras variables ambientales que existan entre los sitios. Así, el modelo establece que el efecto de competencia entre las plantas es general y no depende o varía de un sitio a otro.

El diseño experimental apropiado para este modelo es un diseño factorial con un factor “Competidor” con dos niveles: presente a densidad natural o removido y otro factor “sitio”, con tantos niveles como lugares en que se repita el experimento. El modelo explicativo en este caso requiere que la interacción entre los dos factores NO sea significativa. Si la interacción es significativa, entonces la competencia entre las especies varía, por alguna razón entre los sitios.

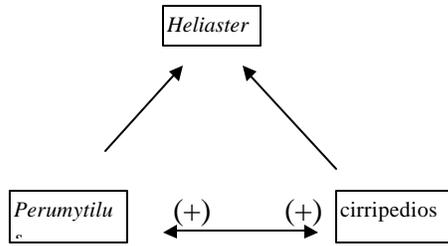
2. Se realizó la observación que la abundancia de la especie *Perumytilus*, el chorito intermareal, es menor cuando la abundancia del depredador *Heliaster helianthus*, sol de mar es mayor. Además, se ha observado que este efecto es menos pronunciado cuando en el hábitat hay *Chthamalus*, cirripedio intermareal.

El modelo explicativo que proponemos es el siguiente: Postulamos que al estar presente *Chthamalus* en el hábitat, los soles de mar consumen tanto *Chthamalus* como *Perumytilus*, disminuyendo así la presión de depredación sobre los choritos. El fenómeno propuesto se denomina en ecología “mutualismo aparente”, por cuanto la presencia de una especie presa secundaria beneficia una especie presa primaria, en forma indirecta, a través de contribuir a la saciación del depredador. Sin embargo, nuevamente existen explicaciones alternativas. Por ejemplo, los cirripedios pueden tener efecto positivo sobre los choritos, independiente de la presencia de depredadores. Los cirripedios podrían aumentar la tasa de asentamiento de choritos.

El modelo (A) que sugerimos es el siguiente:



El modelo alternativo (B) considera que el efecto de cirripedios sobre choritos es un efecto directo, independiente de depredación:

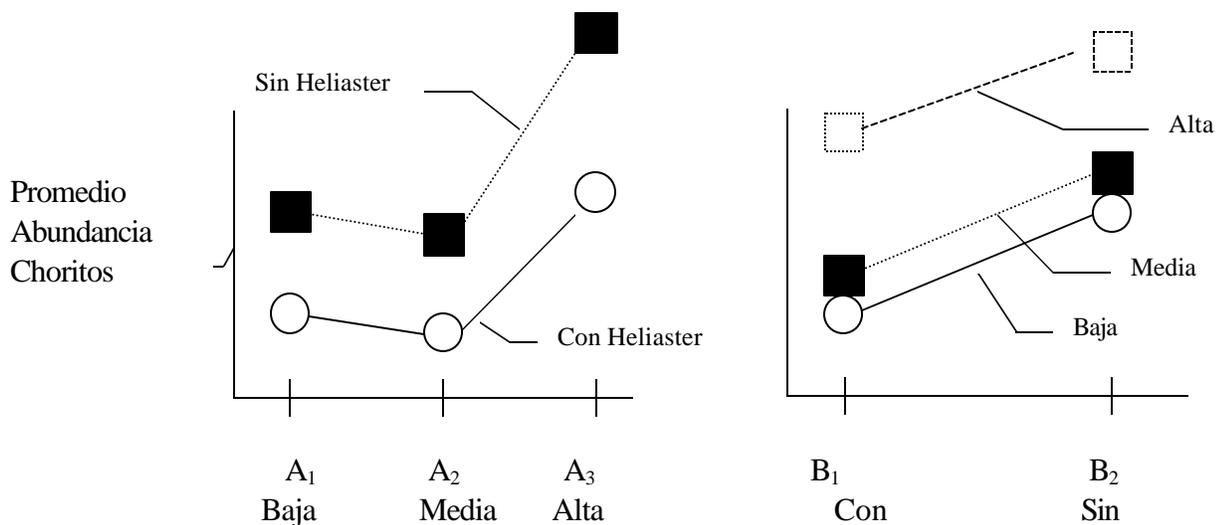


Para evaluar nuestro modelo, realizamos un experimento en el cual manipulamos la presencia de depredadores (*Heliaster* a densidades naturales o removidos manualmente) y la densidad de cirripedios, creando tres niveles de densidad (baja, media y alta) que abarcan el nivel de densidad de cirripedios observados en el ambiente. De esta manera, ambos factores, *Heliaster* y cirripedios, son fijos.

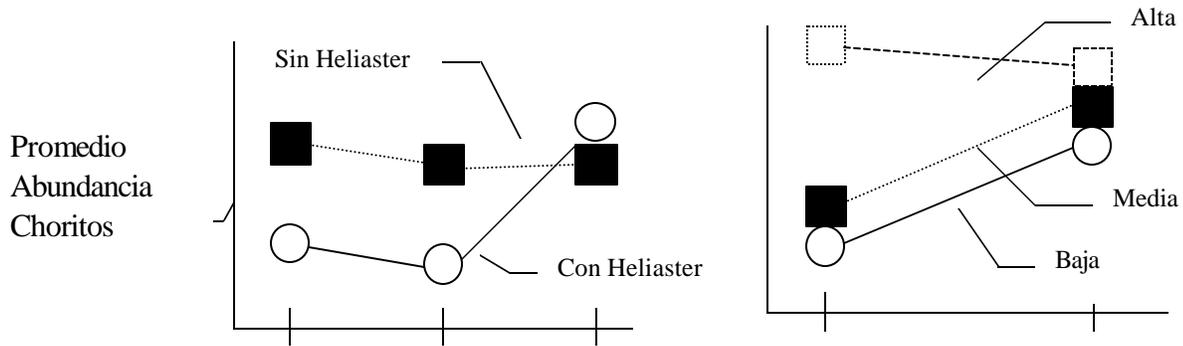
La predicción es que al remover depredadores aumentará la abundancia de choritos. Pero más importante, predcimos que la densidad de cirripedios tendrá un efecto sobre choritos, solamente cuando hay depredadores presentes. Esto implica que nuestro modelo requiere que la interacción entre los tratamientos sea *significativa*.

Resultados posibles de este experimento en gráficos de interacción:

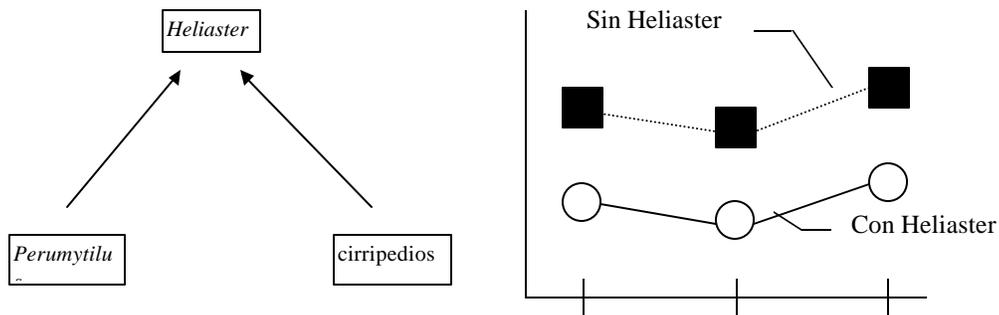
Resultado 1: No consistente con nuestro modelo. Alta densidad de cirripedios aumenta la abundancia de choritos. Los depredadores disminuyen la abundancia de choritos. Sin embargo, el efecto de los cirripedios es independiente de la presencia o no de depredadores, contrario a nuestro modelo explicativo. La diferencia entre los niveles del factor B (depredador) es igual para cada nivel del factor A (cirripedios). Asimismo, ya que el diseño es ortogonal, las diferencias entre la abundancia de choritos en presencia o ausencia de depredadores es la misma para cada nivel de densidad de cirripedios.



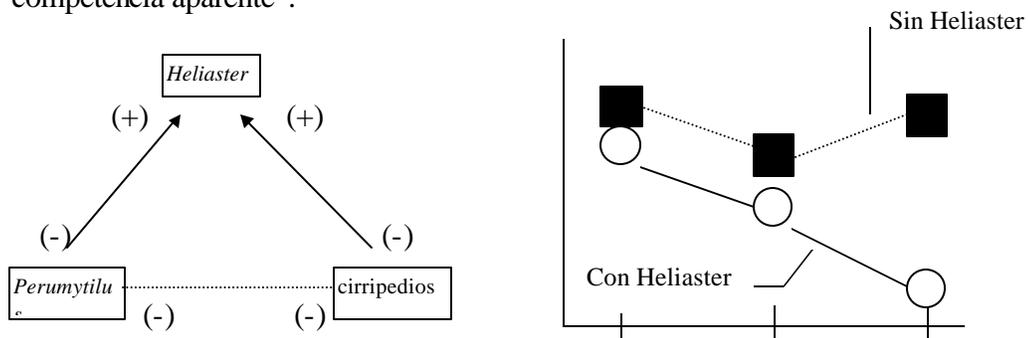
Resultado 2: Consistente con nuestro modelo: Los cirripedios tienen un efecto sobre la abundancia de choritos, pero solamente cuando los depredadores están presentes. No existe un efecto directo de cirripedios sobre choritos, pero si existe un efecto positivo indirecto.



Otros resultados posibles son 3) que no exista interacción (factores independientes), ni tampoco un efecto de cirripedios, pero sí efecto de depredación. En este caso, los modelos explicativos A y B estarían equivocados y en el único modelo apropiado es:



Un cuarto resultado posible es que exista interacción, pero que efectivamente la presencia de cirripedios disminuya la abundancia choritos en presencia de depredadores. Este resultado estaría en desacuerdo con nuestras observaciones iniciales y deberíamos probablemente asegurarnos de controlar otros factores. En este caso, el modelo explicativo más apropiado sería el “competencia aparente”:



En una tabla de ANDEVA los resultados anteriores son:

Fuente de Variación	g.l.	Result. 1	Result. 2	Result. 3	Result. 4
<i>Heliaster</i> (A)	a-1	*		*	
cirripedios (B)	b-1	*		ns	
Interacción (A*B)	(a-1)(b-1)	ns	*	ns	*
Error (dentro de grupos)	ab (n-1)				

2. Eficiencia y costo-efectividad de un Diseño Factorial

La otra gran ventaja de usar diseños factoriales, además de la obvia ganancia en la información entregada por la interacción entre factores, es la ganancia en eficiencia y poder de un diseño factorial por sobre ANDEVA de una vía separados.

Un ejemplo puede aclarar esto. Supongamos que diseñamos un experimento para ver si la administración de un shock de temperatura produce una mayor concentración de proteínas reparadoras en las células de una especie de bivalvo, en comparación a controles y controles de procedimiento apropiados. La hipótesis ha sido propuesta, separadamente, para hábitats hidrotermales superficiales y hábitats hidrotermales profundos. Puesto que la hipótesis ha sido planteada en forma separada para los dos hábitats (B_1 y B_2), es perfectamente razonable realizar dos experimentos separados, uno para cada hábitat, y analizar los datos de acuerdo a dos ANDEVA de una vía separados. El análisis en cada hábitat consiste en tres niveles del tratamiento temperatura (temperatura ambiental, temp. alta, manipulación sin cambio de temp.), con nueve réplicas independientes.

Ahora, puesto que la misma hipótesis está siendo examinada en cada hábitat, el experimento podría bien ser considerado como un diseño factorial de dos factores, con hábitat como el segundo factor.

Asumamos que no hay Interacciones, de manera que podamos realizar tests sobre los factores principales y comprar con los ANDEVA de una vía.

		Exp. 1 Hábitat B ₁	Exp. 2 Hábitat B ₂	
Fuente de Variación	g.l.	CME	g.l.	CME
Temp. (A)	2	$\sigma_e^2 + 9T_A^2$	2	$\sigma_e^2 + 9T_A^2$
Residuo	24	σ_e^2	24	σ_e^2
TOTAL	26		26	

Prueba de Hipótesis para efecto de Temperatura: CM_A/CM_{error} , con 2 y 24 g.l.
Costo es 9 replicas por 3 niveles de A, por 2 hábitats = 54 unidades experimentales

El mismo experimento pero ahora considerado como un diseño factorial:

Fuente Variación	g.l.	CME
A (temp.)	2	$\sigma_e^2 + 18T_A^2$
B (hábitat)	1	$\sigma_e^2 + 27T_B^2$
A * B (interacción)	2	$\sigma_e^2 + 9T_{AB}^2$
Error residual	48	σ_e^2
TOTAL	53	

Prueba de Hipótesis para efecto de Temperatura: CM_A/CM_{error} , con 2 y 48 g.l.
Costo es 9 replicas por 3 niveles de A, por 2 hábitats = 54 unidades experimentales

El mismo diseño con solamente cinco réplicas para cada hábitat nos da un diseño comparable al de diseño de análisis de una vía:

Fuente Variación	g.l.	CME
A (temp.)	2	$\sigma_e^2 + 10T_A^2$
B (hábitat)	1	$\sigma_e^2 + 15T_B^2$

A * B (interacción)	2	$\sigma_e^2 + 5T_{AB}^2$
Error residual	24	σ_e^2
TOTAL	29	

Ahora, la prueba de hipótesis de temperatura también tiene 2 y 24 grados de libertad, como en los ANDEVA separados, pero el costo es solamente: 3 niveles de A x 2 niveles de B x 5 replicas = 30 unidades experimentales.

III. CME EN MODELOS FIJOS, ALEATORIOS Y MIXTOS

Es importante entender que para poder completar cualquier análisis de varianza factorial, es necesario primero determinar si los factores bajo estudio son fijos o aleatorios. Este punto es crucial, pues determina que varianzas hemos estimado o calculado al calcular los cuadrados medios en la ANDEVA.

El caso más simple es cuando ambos factores son fijos:

Para el caso de ANDEVA de dos vías **Modelo I (dos factores fijos)**, la Tabla de ANDEVA se ve así.

Fuente de Variación	g.l.	SC	CM	CME	F
Tratamiento A (T1)	a-1	SC _{T1}	SC _{T1} /a-1	$\sigma_e + nbT_A^2$	$\frac{CM_{T1}}{CM_{Err}}$
Tratamiento B (T2)	b-1	SC _{T2}	SC _{T2} /b-1	$\sigma_e + naT_B^2$	$\frac{CM_{T2}}{CM_{Err}}$
Interacción (A*B)	(a-1)(b-1)	SC _{Int}	SC _{Int} /(a-1)(b-1)	$\sigma_e + nT_{AB}^2$	$\frac{CM_{Int}}{CM_{Err}}$
Error (dentro de grupos)	ab (n-1)	SC _{Err}	SC _{Err} /ab(n-1)	σ_e	

Las pruebas de hipótesis para el factor A, factor B y la interacción usan el error residual como fuente de variación dentro de grupos:

$$F_A = CM_A / CM_{error}$$

$$F_B = CM_B / CM_{error}$$

$$F_{AB} = CM_{AB} / CM_{error}$$

Cuando los tratamientos son aleatorios (A y B aleatorios), entonces la tabla de ANDEVA se ve así.

Fuente Variación	de	g.l.	SC	CM	CME
Tratamiento A		a-1	SC _{T1}	SC _{T1} /a-1	$\sigma_e + n\sigma_{AB}^2 + nb\sigma_A^2$
Tratamiento B		b-1	SC _{T2}	SC _{T2} /b-1	$\sigma_e + n\sigma_{AB}^2 + na\sigma_B^2$
Interacción (A*B)		(a-1)(b-1)	SC _{Int}	SC _{Int} /(a-1)(b-1)	$\sigma_e + n\sigma_{AB}^2$
Error (dentro de grupos)		ab (n-1)	SC _{Err}	SC _{Err} /ab(n-1)	σ_e

En este caso las pruebas de hipótesis son:

$$F_A = CM_A/CM_{AB}$$

$$F_B = CM_B/CM_{AB}$$

$$F_{AB} = CM_{AB}/CM_{error}$$

Cuando un factor es fijo (A) y el otro es aleatorio (B), entonces tenemos lo siguiente:

Fuente Variación	de	g.l.	SC	CM	CME
Tratamiento A (T1)		a-1	SC _{T1}	SC _{T1} /a-1	$\sigma_e + n\sigma_{AB}^2 + nb T_A^2$
Tratamiento B (T2)		b-1	SC _{T2}	SC _{T2} /b-1	$\sigma_e + na\sigma_B^2$
Interacción (A*B)		(a-1)(b-1)	SC _{Int}	SC _{Int} /(a-1)(b-1)	$\sigma_e + n\sigma_{AB}^2$
Error (dentro de grupos)		ab (n-1)	SC _{Err}	SC _{Err} /ab(n-1)	σ_e

Las pruebas de hipótesis entonces son:

$$F_A = CM_A/CM_{AB}$$

$$F_B = CM_B/CM_{error}$$

$$F_{AB} = CM_{AB}/CM_{error}$$

