Capítulo V

Introducción a la Estadística Inferencial (1)

Poblaciones, Muestreos e Inferencia

Se puede definir la Inferencia como el proceso de hacer aseveraciones lógicas sobre lo desconocido en base a alguna evidencia comprobable. En estadística, aquello desconocido sobre lo que se trata de sacar conclusiones se llama **población** y la evidencia que se está buscando son los datos obtenidos a partir de **muestras** tomadas de esa población.

Se puede definir una población como una colección de organismos, cosas o eventos, que constituyen el conjunto completo de individuos (o cosas o eventos) que se estudiar. Si por ejemplo, se desea hablar acerca de gatos, se puede definir la población de gatos como todos los gatos del mundo (si se desea hablar de los gatos en el sentido más amplio), todos los gatos en las Islas Galápagos (si los gatos de Galápagos son el grupo de nuestro interés), o todos los gatos negros con ojos verdes que viven en Puerto Ayora (si son los únicos organismos que nos interesan). Como puede verse, una población puede consistir de muchos individuos, unos pocos, o tan sólo uno.

En estadística es muy importante ser capaz de definir exactamente cuál es la población de interés. Esto es especial cuando se diseñan experimentos, cuando se analizan los datos, y más importante aún, cuando se obtienen las conclusiones. Más adelante se verá la verdad de esta afirmación.

Cuando una población es muy pequeña, puede resultar práctico obtener datos de cada individuo en la población. En tales casos se pueden señalar aspectos acerca de la población (por ej. media, varianza) con absoluta seguridad (al menos con la seguridad que la precisión de nuestras mediciones lo permita). Si se ha contado cada individuo en una población, entonces se puede decir, por ejemplo, exactamente cuántos individuos forman la población, la media de su peso, y el radio entre hembras y machos.

Desafortunadamente, en Biología muy rara vez existe la oportunidad de tener medidas de una población entera. Normalmente la población en cuestión es demasiado grande como para permitir un muestreo exhaustivo. En estos casos se está obligado a depender de datos tomados a partir de una fracción del grupo total. Se hace referencia a esta fracción como **muestra**. A partir de esta muestra de datos, se hacen deducciones (inferencias) acerca de las características de la población. Este es el proceso de **Inferencia Estadística**.

Una de las preocupaciones más importantes en diseño experimental es obtener muestras apropiadas (es decir, verdaderamente representativas). Hay libros y cursos enteros dedicados al estudio de cómo hacer diseños experimentales apropiados. Algo fundamental para en diseño experimental es saber cómo tomar muestras estadísticamente válidas y útiles. Es decir, conseguir muestras representativas, no sesgadas y que reúnan todos los requerimientos de las pruebas que se van a utilizar para analizarlos. A continuación se discutirá cómo conseguir esto.

La habilidad para diseñar buenos experimentos, estadísticamente correctos, no es algo fácil de aprender. Este libro difícilmente va a convertir a nadie en un mago, pero se espera que lo que constituya una ayuda para entender el proceso de diseñar experimentos y cómo evitar errores comunes al hacerlo. Son pocos los investigadores (estudiantes, profesores, científicos profesionales, etc.) que son realmente buenos para diseño experimental. No importa quién sea, o cuánta experiencia tenga, es siempre una buena idea que otras personas revisen sus experimentos antes de que comience. Pequeños errores en el diseño pueden ser suficientemente grandes como para reducir el valor de un experimento. Un error de diseño aparentemente pequeño pero fundamental puede invalidar potencialmente aún el más elaborado (y caro) de los experimentos.

¿Qué son Muestras al Azar?

Como se dijo, es muy importante que una muestra sea representativa de la población de donde salió. Hay muchas maneras de obtener muestras, muchas de ellas malas y unas pocas buenas. Para sacar conclusiones basadas en inferencia estadística la mayoría de pruebas requiere que los individuos (o réplicas) que componen la muestra sean de alguna manera seleccionados al **azar**. El muestreo al azar requiere que cada individuo tenga la misma oportunidad de ser escogido. Más aún, la selección de un individuo no debe afectar la probabilidad de seleccionar cualquier otro. Por ejemplo, la selección de un individuo no debe de ninguna manera influir (o ser influida por) la probabilidad de escoger a su vecino.

Lo anterior nos regresa al tema de qué constituye una población. Es muy importante que se identifique la población *antes* de iniciar el muestreo. Si por ejemplo, se desea muestrear árboles en Santa Cruz (en forma azarosa, por supuesto) y más tarde decidir que se quiere sacar conclusiones acerca de todos los árboles en Galápagos, se debe encontrar un diseño de muestreo que nos permita hacerlo. Cuando se decide muestrear sólo árboles en Santa Cruz, automáticamente las conclusiones que se pueden obtener se ven limitadas a los árboles de Santa Cruz. ¿Por qué? Porque la única población de árboles para los cuales se tiene una muestra al azar (representativa) es la de Santa Cruz. La misma limitación existe para todos los diseños experimentales. Las limitaciones de los diseños experimentales estarán siempre restringidas a la naturaleza y extensión de las conclusiones. Por lo tanto, al diseñar un experimento se debe primero decidir la amplitud de las conclusiones que se desea hacer, y entonces diseñar el experimento con estas bases.

No todos los diseños experimentales son (o deben ser) completamente al azar en todos los aspectos. En muchas ocasiones los elementos de muestreo no aleatorios son necesarios en el diseño para contestar cierto tipo de preguntas. Por ejemplo, en un estudio sobre la efectividad relativa de veneno sobre cabras machos y hembras, se puede seleccionar intencionalmente un número fijo de hembras y de machos de la población, en vez de escoger los animales completamente al azar. El elemento no aleatorio aquí es el sexo, ya que el estudio más efectivo tendrá un número igual de machos y hembras. Fíjese que mientras escogemos el número de individuos de cada sexo, cada individuo de cada sexo es todavía escogido al azar. Un experimento puede tener muchos niveles (o elementos) no aleatorios. Sin embargo, todos los experimentos que van a generar datos para hacer un análisis inferencial posterior, deben tener al menos algún elemento o elementos de azarosidad. Desafortunadamente no se van a estudiar aquí los muchas tipos posibles de diseño experimental.

¿Qué es Pseudo Replicación?

La pseudo replicación recibe su nombre de la creencia de que ciertos métodos de muestreo no aleatorios resultan en muestras estadísticamente independientes (réplicas) cuando realmente no lo son, pues hay **pseudo-réplicas** en vez de verdaderas réplicas. Como ejemplo, considérese un investigador que cada día debe determinar la calidad de las tortas producidas en una pastelería. La población diaria es el número total de tortas hechas ese día. En vez de seleccionar y probar varias tortas individuales, este investigador decide seleccionar solo una torta, pero la corta en varios pedazos y prueba cada uno. Esto *es* pseudo replicación.

Probar la calidad de los pedazos de una torta nos da información solo de la variación en calidad en esa torta individual y no nos dice nada acerca de la variación en la calidad entre tortas, que era lo que interesaba aquí. Considere las conclusiones falsas a las que este diseño puede llevar. Si se asume por ejemplo, que en realidad 50% de las tortas tiene una calidad más baja de lo esperado, este investigador tiene una probabilidad del 50% de perder una torta de buena calidad. ¡También, puede ser que la torta elegida sea de buena calidad en cuyo caso la conclusión sería que *todas* las tortas del día son de buena calidad y uniforme!

Desafortunadamente, en experimentos biológicos la pseudo-replicación no es siempre tan obvia como lo fue en el ejemplo anterior. Hay muchos ejemplos que requieren una consideración muy cuidadosa antes de que se vea por qué existe pseudo-replicación. Para entender mejor esto en Biología, se estudiarán un par de errores de los más comunes.

Un investigador ha sido contratado para estudiar el efecto de la temperatura de incubación sobre tortugas terrestres por eclosionar. Se conoce que hay tres regímenes de temperaturas de incubación (caliente, templado y frío), tres incubadoras y 90 huevos (obtenidos de una variedad de hembras). El investigador decide colocar 30 huevos (escogidos al azar de las varias hembras) en cada incubadora y colocar cada incubadora en cada uno de los regímenes de temperatura.

¿Qué es lo que está mal con este diseño? ¿Dónde y por qué hay pseudo-replicación? El problema está en el uso de un solo incubadora para cada una de las condiciones de incubación. Durante el experimento, todos los huevos en uno de las incubadoras están sujetos no solo a un régimen de temperatura único, sino también a todo el conjunto de condiciones característico de esa incubadora en particular. Por ejemplo, se sabe que cierto hongo, muy difícil de detectar, infecta huevos de tortugas, lo cual resulta en un peso bajo en los que eclosionan. Si el hongo atacara uno de las incubadoras es probable que muchos de los huevos en tal incubadora sean infectados. El resultado sería la eclosión de tortugas de menor tamaño en esa incubadora. Si el investigador no nota la infección micótica, la conclusión para el experimento será que en esa temperatura las tortugas eclosionan con bajo peso. Tal conclusión puede ser totalmente errada. Estadísticamente, para este experimento en lugar de 90 réplicas verdaderas, hay tres. Una réplica por cada temperatura, que es el peso medio de los eclosionados por incubadora. En efecto, sería imposible analizar estos datos debido al diseño experimental tan pobre.

La lección aquí es que un individuo (en el sentido estadístico) debe estar igualmente sujeto a la posibilidad de todas las fuentes de variación (variación que no está bajo el control directo del experimentador). En el ejemplo anterior, la probabilidad del influjo de hongos no está compartida

igualmente por todos los huevos. Si un hongo contamina una incubadora todos los huevos que se encuentran en él tendrán mucha mayor probabilidad de ser afectados. No hay manera de diferenciar entre el influjo del hongo y de la temperatura. La respuesta a este problema puede ser aislar cada huevo de su vecino, ya sea colocándolo en un incubadora independiente (sería lo mejor), o dividiendo un incubadora en compartimientos aislados y colocando un huevo en cada división (solución no tan buena, pero práctica).

En otro ejemplo el objetivo de la investigación es estimar la densidad de árboles de *Scalesia* en un bosque de 1000 ha. Se ha sugerido al investigador encargado que el método más conveniente de muestreo sería ir al bosque, seleccionar un área 'representativa' de 200 x 200 m, dividirla en 100 cuadrantes de 20 x 20 m. y luego determinar la densidad de *Scalesia* en cada cuadrante. Se le ha señalado también, que con esta división resultarían 100 réplicas, lo que daría una buena estimación de la densidad de tales árboles en el bosque.

En este momento se conocen muy bien las causas y consecuencias de la pseudo-replicación y resulta obvio que lo que se ha llamado 'réplicas' son en realidad pseudo-réplicas, porque sólo son subdivisiones de una réplica grande (el cuadrante de 200 x 200). Cada cuadrante no puede ser una réplica tomada al azar de la población total (el bosque) porque ha sido tomada del área restringida del cuadrante. No importa que este cuadrante se haya escogido como 'representativo', ya sea por escoger su localización intencionalmente o seleccionando coordenadas al azar; todos los cuadrantes tomados de él son todavía "pedazos de la misma torta" y no pueden ser usados para estimar la densidad de todo el bosque. Para evitar la pseudo-replicación, en vez de esto, el investigador decide localizar al azar cada uno de los 100 cuadrantes de 20 x 20 m. en todo el bosque.

Inferencia Estadística

Se mencionó que la inferencia estadística es el proceso de sacar conclusiones generalizadas sobre una población (o poblaciones) basadas en datos obtenidos a partir de muestras. Generalmente las inferencias son declaraciones sobre diferencias. Por ejemplo, se puede querer decir que dos clases de individuos son diferentes (en alguna variable) entre ellos, o tal vez decir que una distribución no es al azar, etc. En estadística el tipo de inferencia hecha se establece formalmente, usando dos tipos de declaraciones mutuamente exclusivas, o hipótesis, conocidas como la **Hipótesis Nula** (H_O) y la **Hipótesis Alternativa** (**H_a**). Las dos hipótesis describen dos posibilidades alternativas. Se escoge H_O y Ha para declarar específicamente qué es lo que se está probando.

H_O es generalmente una declaración de no haber diferencia y es la versión de la realidad que se aceptará a menos que se tenga suficiente evidencia para rechazarla y aceptar H_a. Como ejemplo, H_O puede declarar que no hay diferencia entre las medias de dos poblaciones, o que el valor de una variable es independiente de la otra, o que la distribución espacial de una especie dada es al azar.

Por otro lado, H_a es una declaración de diferencia y es la versión de la realidad que debemos aceptar tentativamente si se encuentra suficiente evidencia para rechazar H_o. H_a es la única alternativa para H_o. Por ejemplo, si H_o establece que no hay diferencia, entonces H_a establecerá que hay una diferencia. Otros ejemplos para H_o / H_a son independencia/dependencia y azarosidad/no-azarosidad.

En estadística a menudo se habla acerca de **significancia**. Cuando se dice que algo es significativamente diferente lo que realmente se quiere decir es que se tiene suficiente evidencia en la muestra para rechazar la versión de realidad que llamamos H_o en favor de H_a. Otra manera de ver esto es que a menos que se tenga suficiente evidencia de lo contrario se debe aceptar la versión de realidad llamada H_o .

Es muy importante recordar dos cosas sobre la inferencia estadística. Primero, el no rechazar H_o **NO** es lo mismo que **PROBAR** que H_o es **VERDADERA**. Segundo, aceptar H_a (rechazar H_o) **NO** es lo mismo que **PROBAR** que H_a es **VERDADERA**. ¿Por qué es esto así? Porque todo el proceso de inferencia estadística está basado en tratar de hacer declaraciones sobre muchos individuos en base a apenas unos cuantos.

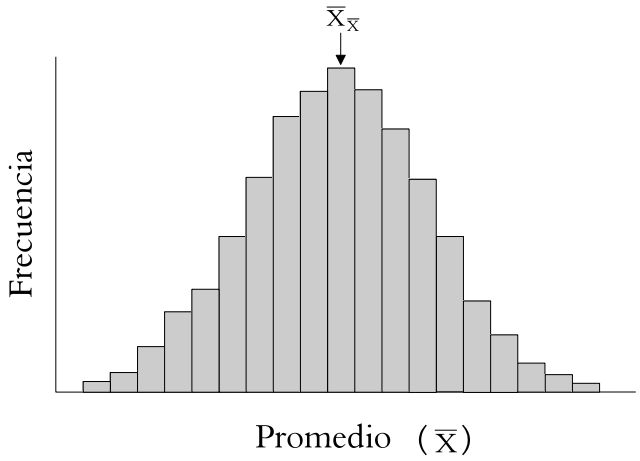
Muestreo al azar significa que siempre es posible que una muestra dada esté compuesta de algunos de los valores más extremos en la población. Por lo tanto, siempre es posible que cualquier ejemplo dado sea muy diferente en realidad de la mayoría de individuos. Las inferencias que se hacen acerca de la población basados en tal muestra podrían estar muy lejos de la realidad. Así, en el proceso de inferencia estadística siempre es posible que cualquier conclusión (inferencia) que se saque en base a una muestra de una población sea errada.

Cuando se habla de tener 'suficiente evidencia para rechazar H_o' lo que realmente se dice es que la probabilidad de error es suficientemente pequeña como para permitir la aceptación de H_a. La razón para que el muestreo al azar sea tan importante en la inferencia estadística, se debe a que permite conocer exactamente cuál es la probabilidad de rechazar H_o incorrectamente. Se tratará esto más adelante.

Distribuciones de Muestreo

Así como los datos de una variable dada tienen una distribución (normal, uniforme, binomial, etc.) los estadísticos de muestras (media, varianza, etc.) también tienen sus propias distribuciones. Imagínese por ejemplo, que a partir de una población de plantas muy grande se escoge al azar repetidamente cierto número de semillas y se pesan. De cada una de estas muestras se calcula la media del peso de las semillas. Ahora se dibuja un histograma de estas medias como se ve en el Gráfico 5.1. Lo que se ve es que, en el caso de las medias, a medida que el número de muestras aumenta, la distribución de las medias tenderá a parecerse a una **distribución normal**. Esto es cierto a pesar de la distribución de los pesos de las semillas individuales. Este fenómeno, que las muestras tiendan a una distribución normal, es muy importante en estadística y se conoce como el **Teorema de Límite Central**.

Gráfico 5.1



Se conocen un número de características de la distribución de las medias. La primera es que el valor de la media de las medias es también una estimación no sesgada de :. El valor de la media de las medias es igual a:

$$\overline{X}_{\overline{X}} = \frac{\sum \overline{X}_i}{r} \qquad (5.1)$$

Donde $_i$ es la media de la muestra iésima y **r** es el número de replicaciones.

Más aun, la varianza de las medias alrededor de la media de las medias es el **error estándar** (). La fórmula para el error estándar es:

$$S_{\overline{X}} = \frac{s}{\sqrt{n}} \qquad (5.2a)$$

$$S_{\overline{X}} = \sqrt{\frac{s^2}{n}} \qquad (5.2b)$$

Donde **n** es el tamaño de la muestra y **s** es la desviación estándar de la muestra.

Se ve que la varianza en las medias decrece cuando el tamaño de las muestras crece. Esto se puede entender mejor si se considera que la precisión de cada estimativo de la media de la muestra se incrementará al tiempo que el tamaño de la muestra crezca. Por ejemplo, una muestra que contiene 99% de los individuos de una población de un millón de individuos es mucho más capaz de reflejar la media verdadera de la población que una muestra de un sólo individuo. En esta forma un número de muestras de gran tamaño de la misma población tenderán a ser buenos

estimativos de : y por lo tanto tenderán a ser muy similares (la varianza de los estimativos serán relativamente reducida). Muestras pequeñas tenderán a dar peores estimativos de : y por lo tanto a estar más ampliamente distribuidas en sus valores (la varianza de los estimativos serán relativamente grande).

Como se verá, algunos de estadísticos tienen distribuciones que son suficientemente importantes para ameritar sus propios nombres.

Hipótesis de Una Muestra sobre la Media

Es tiempo de considerar una de las pruebas inferenciales más simples en estadística. La forma general de esta prueba, sin embargo, se repite en un rango amplio de procedimientos más complicados.

En estadística a menudo se quiere saber si la media de una población es diferente de algún valor hipotético. Dado que es raro poder estudiar cada individuo en una población, es necesario saber cómo hacer inferencias sobre los valores de la media de una población (:) basados en la media de una muestra ().

Como ejemplo, se desea conocer si el peso medio ganado por los pinzones mientras comen arroz cubierto con veneno para ratas es diferente de 0. Para analizar los datos se debe primero decidir las dos hipótesis teóricas: H_o y H_a. Después de considerarlo cuidadosamente se decide que:

H_o: El peso medio ganado por pinzones es igual a 0. (: = 0)

H_a: El peso medio ganado por pinzones es diferente de 0. (: <> 0)

La prueba que se empleará es la llamada **prueba t** desarrollada por W.S. Gosset, quien publicó sus teorías bajo el nombre de ‘Student’, por esto se hace referencia a esta prueba como “**t de Student”**. La fórmula básica para esta prueba es:

$$t = \frac{\overline{X} - \mu}{S_{\overline{X}}} \qquad (5.3)$$

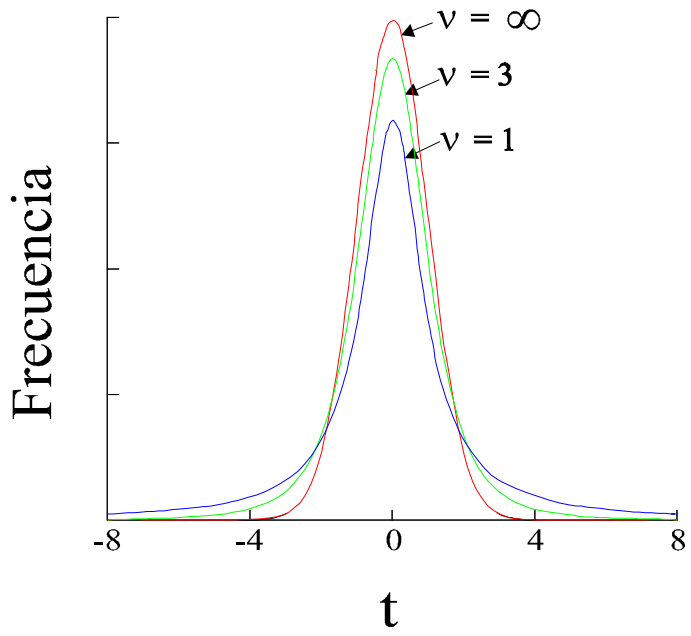
Donde : es el valor teórico que se desea probar (0 en nuestro ejemplo).

Como ocurre con las medias de las muestras, los valores de **t** también tienen su propia distribución, que es conocida como **distribución t**. La forma de tal distribución depende del tamaño de la muestra **n**, como se ve en el Gráfico 5.2.

Como se ve, la distribución se hace más y más angosta a medida que el tamaño de la muestra (también conocido como grado de libertad) incrementa. Cuando el tamaño de la muestra

es infinitamente grande (un imposible, pero los estadísticos rara vez se preocupan por cosas tan triviales) la distribución t es idéntica a la distribución normal.

Gráfico 5.2



La media de la distribución t es igual a 0 y corresponde a = :. Los valores a la derecha y a la izquierda de : (valores positivos y negativos de **t**) representan valores de **t** donde la media es menor o mayor que :, respectivamente. Como se puede ver en el Gráfico 5.2, siempre es posible que la magnitud de la diferencia entre la media y : sea menor que mayor. Esto puede ser interpretado considerando que si H_o es cierta (la media de la población = la media teórica) entonces es más difícil que ocurran grandes diferencias entre las medias de las muestras y : que pequeñas diferencias, basados puramente en el azar. Por otro lado, si la media de la población y : no son iguales, los valores de t tenderán a ser mucho más diferentes de 0. En algún momento, cuando el valor de t es suficientemente diferente de 0, se debe rechazar H_o y aceptar que en efecto, hay una diferencia (H_a).

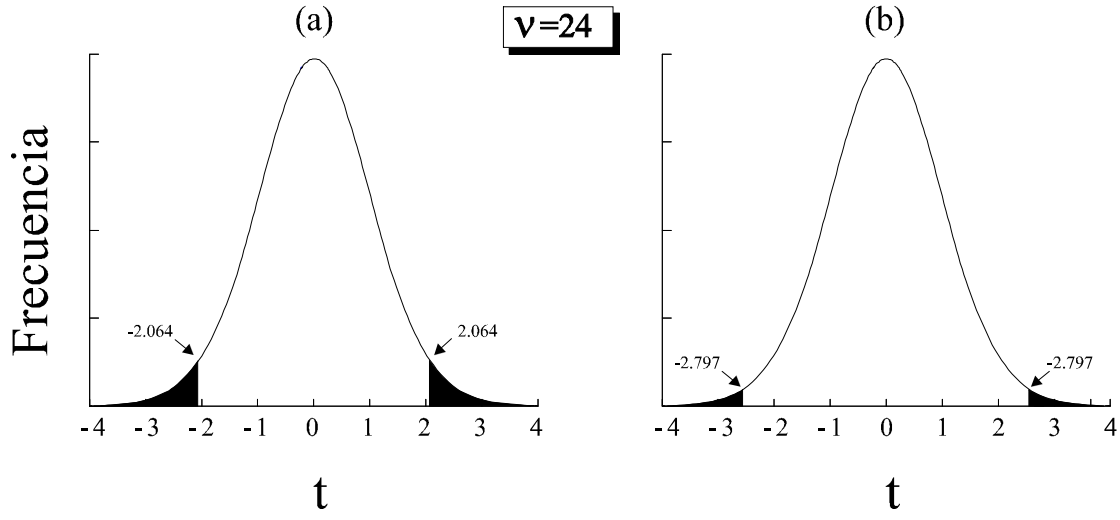
¿Cómo se sabe cuándo rechazar H_o? Para responder esto se deben considerar las consecuencias de aceptar o rechazar H_o. Recuerde que por el hecho de que siempre se trabaja con evidencia que provee muestras al azar, existe el riesgo de que cuando se rechaza H_o se esté cometiendo un error. Igualmente, cuando no se rechaza H_o se corre el riesgo de que H_a sea en verdad válida. Nos referimos a estas fallas como **errores** y cada uno tiene un nombre especial.
El rechazo de H_o cuando H_o es verdadero (no hay diferencia) se conoce como **Error Tipo I**.
El no rechazo de H_o cuando H_a es verdadero (si hay una diferencia) se conoce como **Error Tipo II**.

La probabilidad de cometer un error tipo I es también conocida como **nivel de significancia** de la prueba y se le da el símbolo **α**. A la probabilidad de cometer un error tipo II se le da el símbolo **ß**.

Esto se puede interpretar en el Gráfico 5.3. El área bajo la curva de una distribución t es igual al 100% de los valores de t. El 5% de los valores más extremos posibles se muestran en las partes sombreadas del Gráfico 5.3a. Estadísticamente se espera que si no hay diferencia entre la media de la población y el valor teórico, puramente por azar 5% de nuestras muestras deberían resultar en un valor de t en algún lugar de esos extremos. Cuando se escoge algún nivel de significancia, lo que se quiere decir es que cuando un valor de t es suficientemente extremo (la probabilidad de que este valor ocurra por puro azar cuando H_o es verdadero es suficientemente pequeño) se rechaza la posibilidad de H_o en favor de H_a. Por lo tanto cuando se rechaza H_o siempre se corre el riesgo de estar errados y este riesgo es exactamente igual a α (nivel de significancia). El escoger $\alpha = 5\%$ es completamente arbitrario. No hay razón para que no podamos escoger $\alpha = 10\%$, o 1% (Gráfico 5.3b), etc. Por convención, usamos 5% como el nivel de probabilidad mínimo significativo porque se dice que 5% es un nivel aceptable de riesgo. Mientras más pequeño sea el valor de α, más pequeño será este riesgo.

Contrariamente, la probabilidad de cometer un error tipo II (ß) no es generalmente conocida o especificada. En otras palabras, si se rechaza H_o no se sabrá cuál es la probabilidad de que H_a sea verdadera. Se sabe sin embargo, que para una muestra de un tamaño dado (n), α está inversamente relacionada con ß. Así, mientras más pequeño sea α, mayor es la probabilidad de cometer un error de tipo II. También se sabe que ß decrece a medida que n crece: mientras mayor es el tamaño de la muestra menor es la probabilidad de que ocurra un error tipo II.

Gráfico 5.3



Volvamos ahora al ejemplo de los pinzones. Se tienen los siguientes datos de 12 individuos. Cada individuo fue pesado al comienzo del experimento, alimentado por un mes con arroz con veneno

para ratas, y luego pesado otra vez. Los siguientes datos presentan las diferencias en peso entre la primera y la segunda medición.

Tabla 5.1

Cambio de peso (g)		
1.7	-1.2	H_o = : = 0
0.7	-0.9	
-0.4	-1.8	H_a = : <> 0
-1.8	-1.4	
0.2	-1.8	n = 12
0.9	-2.0	

Media del cambio de peso $(\overline{X})$ = -0.65 g
Varianza del cambio (s^2) = 1.568 g^2
Error estándar para el cambio $(S_{\overline{X}})$ = 0.36 g

Grados de Libertad (ν) = n-1 = 12-1 = **11**

Para decidir si rechazar H_o o no, se debe encontrar qué valor de t es suficientemente extremo como para merecer este rechazo. Si se escoge un α = 5% (P = 0.05) se puede mirar el valor de t en una Tabla de valores críticos para el estadístico t, que justamente iguale esta probabilidad. La forma general para un estadístico t crítico es: $\mathbf{t_{\alpha(2),\nu}}$ donde α es el nivel de significancia, *v* son los grados de libertad, y **(2)** indica que esta será una prueba de dos colas.

Se ve en la Tabla 5.1 que $t_{05(2),11}$ = 2.201. La prueba trata de ver la posibilidad de que el valor de la media sea menor o mayor que 0. Por lo tanto, los dos valores extremos, positivo y negativo, de la distribución son de interés en este caso. Este tipo de prueba se conoce entonces como **de dos colas** porque las dos **colas** de la distribución son de interés. En el caso de una prueba de dos colas realmente se tienen dos valores críticos: uno positivo y uno negativo. Las Tablas de valores críticos sólo presentan los valores positivos. Los valores críticos negativos tienen la misma magnitud, pero con signo negativo.

Para decidir si se debe rechazar H_o se debe comparar el valor de la prueba t con los valores críticos. Si el valor absoluto del t de la muestra es mayor que o igual al valor crítico positivo de t, entonces se rechaza H_o en favor de H_a. En el ejemplo presente:

$$t = -1.81 t_{05(2),11} = 2.201$$

Como 2.201 $\geq$ -1.81 no se rechaza H_o. El Gráfico 5.4 es una representación gráfica de esta comparación.

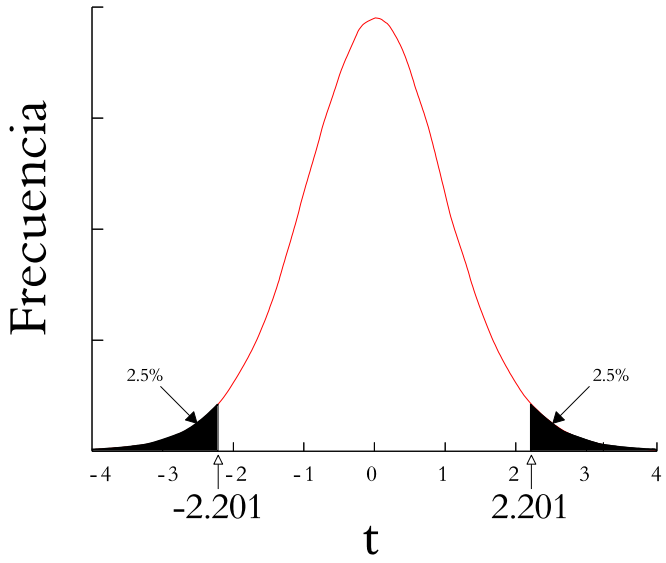
Por lo tanto se concluye que hay insuficiente evidencia para apoyar la hipótesis de que el arroz con veneno para ratas (de una marca particular, en una concentración particular, en un período de tiempo dado) afecta el peso de los pinzones.

La prueba anterior puede ser analizada en términos de una prueba de una cola usando otra hipótesis nula y alternativa. Las pruebas de una cola son casos especiales de la prueba de dos colas que se utiliza cuando hay razones *a priori* para decir que sólo interesan las diferencias en una dirección. En el experimento con el veneno para ratas, por ejemplo, no sería sorprendente que las aves ganen peso si el veneno no tiene efecto ya que ellas estarían recibiendo una cantidad suficiente de arroz. Sin embargo, una prueba más apropiada sería:

$$H_o \geq 0 \qquad H_a < 0$$

Ahora se desea conocer si la media del cambio de peso es significativamente menor que 0, es decir se está interesado en aquellos valores de t que son grandes y negativos. El área crítica de 5% de la distribución t es solo la del extremo izquierdo (véase el Gráfico 5.5). Como ahora se está interesado solamente en una de las colas de la distribución, se necesita duplicar el área usada en el Gráfico 5.4 para poder realmente tomar en cuenta el 5% del área bajo la curva. En este caso la magnitud del valor crítico de t se hace menor.

Gráfico 5.4



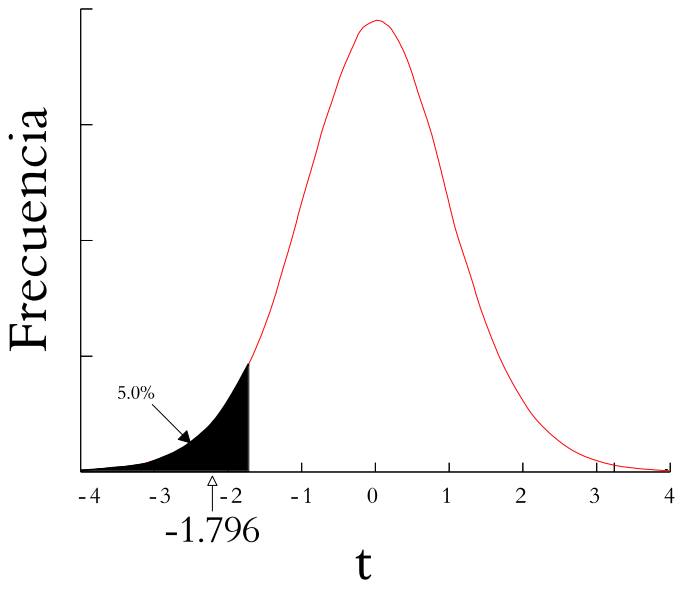
En la Tabla de Valores Críticos de t (página 136) se puede ver que $t_{0.05(1),11} = 1.796$. En este caso el valor t de la muestra analizada es mayor que el valor crítico y por lo tanto se rechaza H_o. Nótese que la prueba de una cola es una prueba más poderosa. En ella, es más fácil (se necesita un valor de t menor para rechazar H_o) mostrar una diferencia significativa con el mismo nivel de probabilidad (de un error tipo I).

Resulta obvio preguntar si una prueba de una cola es más poderosa que una de dos, ¿por qué no se usa siempre la de una cola? La respuesta es que **solo** se puede usar la prueba de una cola cuando se tienen razones para estar interesados en una sola cola de la distribución. En el experimento motivo de este análisis, sería muy razonable decir que el único efecto relacionado con

el peso es la pérdida del mismo, ya que una ganancia sería signo de una dieta saludable. Si no se tiene una razón *a priori* tan clara para usar una prueba de una cola, se debe usar una de dos.

Una nota final acerca de pruebas de una y dos colas: Existe un número de pruebas que por su naturaleza pueden ser de una o dos colas. Para estas pruebas no hay posibilidades de escoger el número de colas. La descripción de estas pruebas le dirá cual debe usar.

Gráfico 5.5



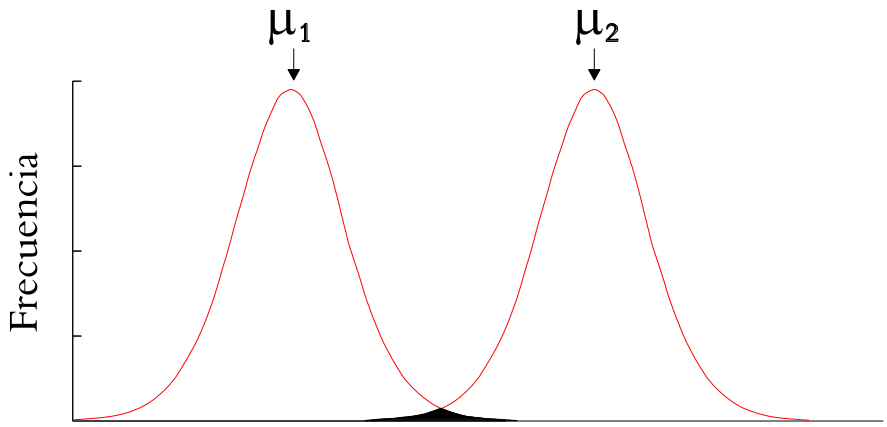
Hipótesis de Dos Muestras sobre la Media

Una de las pruebas bioestadísticas más utilizadas es la comparación de dos muestras para inferir si representan la misma o dos poblaciones distintas. Se puede preguntar también si las medias de dos poblaciones son diferentes. En términos de la hipótesis que se está probando:

H_o: $\mu_1 = \mu_2$ y H_a: $\mu_1 <> \mu_2$

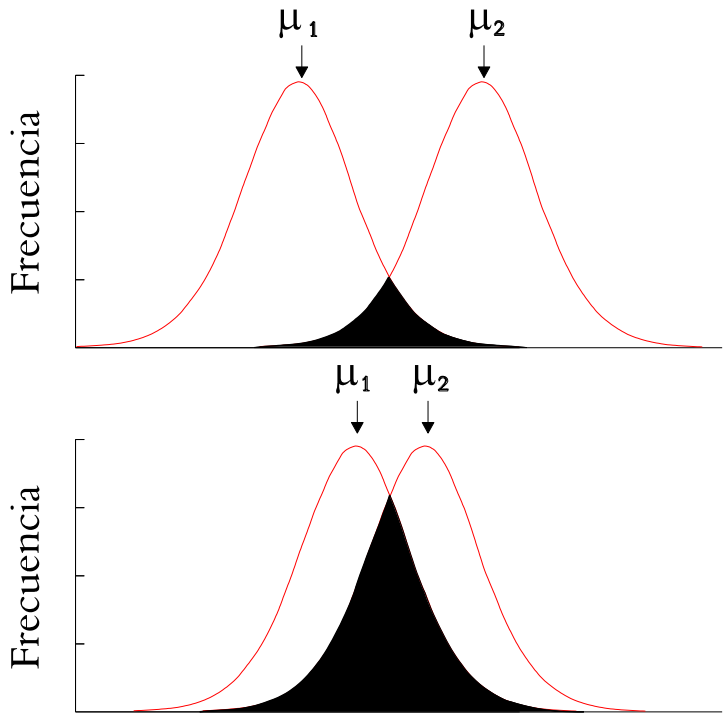
Como en la prueba de una cola, si se muestrean cada una de las dos poblaciones muchas veces y se calculan las medias para cada muestra, se verá que la distribución de las medias de las dos muestras tenderán cada una hacia la distribución normal. Si las medias son muy diferentes las distribuciones podrían lucir como el Gráfico 5.6.

Gráfico 5.6



Nótese que hay muy poca sobreposición entre las dos distribuciones. Así, el rango de valores en común entre las dos distribuciones es pequeño. Si el tamaño de la muestra es moderadamente grande, habrá poca duda de que las medias de las dos poblaciones son diferentes. Sin embargo, a medida que la diferencia de las medias decrece y/o la cantidad de sobreposición entre las distribuciones aumenta (véase el Gráfico 5.7), cada vez será menos obvio que las medias son diferentes.

Gráfico 5.7



Los mismos datos pueden ser representados como las diferencias entre las dos medias:

$$\overline{X}_{1-2} = \overline{X}_1 - \overline{X}_2 \qquad (5.4)$$

Así como con $_1$, $_2$ y $_{1-2}$ también tiende a distribuirse normalmente. Considerado de esta manera, la prueba de dos muestras es análoga a la prueba de una muestra, donde la hipótesis nula es que la diferencia entre las medias no es significativamente diferente de 0. Todos los componentes de las pruebas t de dos muestras tienen análogos en la prueba de una muestra.

Como ejemplo, se considerará el problema de determinar si las medias del peso de las ratas capturadas en dos diferentes zonas en la isla Pinzón son diferentes. Los datos son los siguientes:

Tabla 5.2

Pesos (g)

Area A	Area B
41, 25, 38	52, 57
34, 31, 30	57, 56
33, 37, 40	62, 55
36, 34	55, 64

$n_1 = 11$	$n_2 = 8$
$\nu_1 = 10$	$\nu_2 = 7$
$\overline{X}_1 = 34.45$	$\overline{X}_2 = 57.25$
$\sum X_1 = 379$	$\sum X_2 = 458$
$\sum X_1^2 = 13277$	$\sum X_2^2 = 26328$
$\sum (X_1)^2 = 143641$	$\sum (X_2)^2 = 209764$
$SC_1 = 218.73$	$SC_2 = 107.50$

Hasta ahora no hay nada nuevo. Se conoce todo sobre los cálculos anteriores. Lo mismo que para la prueba t de una muestra, se tiene que calcular la varianza. Sin embargo, dado que se tienen dos grupos de datos se deben combinar los datos de las dos muestras. La varianza resultante se conoce como **varianza conjunta**.

$$s_p^2 = \frac{SC_1 + SC_2}{\nu_1 + \nu_2} \qquad (5.5)$$

$$= \frac{218.73 + 107.50}{10 + 7}$$

$$= \mathbf{19.19}$$

Ahora se calcula el error estándar para los datos conjuntos. Este valor puede ser traducido como la varianza de la distribución de las diferencias entre las medias $(\overline{X}_1 - \overline{X}_2)$.

$$s_{X_1 - X_2} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}} \qquad (5.6)$$

$$= \sqrt{\frac{19.190}{11} + \frac{19.190}{8}}$$

$$= \mathbf{2.0355}$$

Con estos datos se puede calcular el estadístico t para esta prueba.

$$t = \frac{\overline{X}_1 - \overline{X}_2}{S_{\overline{X}_1 - \overline{X}_1}} \qquad (5.6)$$

Los grados de libertad para la prueba son: $\nu = \nu_1 + \nu_2$
$= 10+7$
$= \mathbf{17}$

Como se está interesado en dos posibilidades: $:_1 > :_2$ y $:_1 < :_2$, se llevará a cabo una prueba de dos colas. El valor crítico de t, $t_{0.05(2),17} = 2.110$. Dado que $|-11.199| \gg 2.110$, se rechaza H_o y se acepta H_a. En efecto, $|-11.199|$ es mucho más grande que $t_{0.001(2),17}$ (3.965) y por lo tanto se concluye que con una probabilidad de $P \ll 0.001$ existe una diferencia entre la media del peso de las ratas en las dos zonas.

Una nota final. Se ha utilizado el término grados de libertad (*v*) varias veces. *V* es la manera como se toma en consideración el tamaño de la muestra (n). Pero, ¿por qué entonces se calcula *v* = n - 1, *v* = n - 2, etc? Se pierden grados de libertad cada vez que se tiene que usar un parámetro estimado en los cálculos (típicamente un estimativo de la media). Así, en la prueba t de una muestra, se pierde un grado de libertad cuando se ha tenido que usar la media de la muestra como estimativo de la media de la población. En el caso de la prueba t de dos muestras, se debe estimar dos medias y como resultado se pierden dos grados de libertad. Normalmente la pérdida de uno o dos grados de libertad no es un gran problema (note que *v* debe ser siempre al menos 1). Sin embargo, cuando un experimento incluye muchas fuentes de variación, la carencia de suficientes grados de libertad puede ser un gran problema.

Capítulo VI

Introducción a la Estadística Inferencial (2)

Introducción

En Capítulo V se enseñó cómo hacer inferencias sobre las medias. Específicamente se señaló como usar la prueba t para buscar diferencias entre las medias de algunas poblaciones y un valor teórico, y entre las medias de dos poblaciones diferentes. En ambos casos se usaron las medias de las muestras como los mejores estimativos de las medias desconocidas de las poblaciones.

Intervalos de Confianza

Nunca se puede saber qué tan buena o mala es una estimación de un parámetro de una población a partir de una muestra. Sin embargo, se puede estimar la precisión de las estimaciones calculando los **Intervalos de Confianza** (**IC**).

Recuérdese cuando se está tomando muestras de una población con la media = :, 5% de las muestras tendrán valores de t ($t=(\overline{X}-\mu)/S_{\overline{X}}$) que serán mayor que $t_{0.05(2),\nu}$ o menor que $-t_{0.05(2),\nu}$ (en otras palabras, $|t|>t_{0.05(2),\nu}$). Esto significa que 95% de los valores de t obtenible quedan entre los límites de $-t_{0.05(2),\nu}$ y $t_{0.05(2),\nu}$. Esto se puede escribir como:

$$P\left[-t_{0.05(2),\nu}<\frac{\overline{X}-\mu}{S_{\overline{X}}}<t_{0.05(2),\nu}\right]=0.95 \qquad (6.1a)$$

Arreglando 6.1a nos da:

$$P\left[\overline{X}-t_{0.05(2),\nu}\cdot S_{\overline{X}} \;<\; \mu \;<\; \overline{X}+t_{0.05(2),\nu}\cdot S_{\overline{X}}\right] \;=\; 0.95 \qquad (6.1b)$$

Se puede leer la ecuación 6.1b como, “la probabilidad que la intervalo entre $\overline{X}-t_{0.05(2),\nu}\cdot S_{\overline{X}}$ y $\overline{X}+t_{0.05(2),\nu}\cdot S_{\overline{X}}$ incluye : es 0.95.” La fórmula entre los paréntesis se conoce como el **Intervalo de Confianza** o **Límites de Confianza** y su forma general es:

$$\overline{X}-t_{0.05(2),\nu}\cdot S_{\overline{X}} \;<\; \mu \;<\; \overline{X}+t_{0.05(2),\nu}\cdot S_{\overline{X}} \qquad (6.2)$$

Si se conoce , y *v* entonces se puede decir con una confianza de 95% que : existe dentro del intervalo especificado por fórmula 6.1c. La cantidad se conoce por el **límite de confianza inferior** (L_1) y por el **límite de confianza superior** (L_2).

A partir de la prueba t para una muestra:

$$\begin{aligned} \nu &= 11 \\ \alpha &= 0.05 \\ t_{0.05(2),11} &= 2.201 \\ \overline{X} &= -0.65 \\ S_{\overline{X}} &= 0.36 \\ t_{0.05(2),11} \cdot S_{\overline{X}} &= 0.36 \times 2.201 \\ &= 0.79 \end{aligned}$$

Para el ejemplo que se está siguiendo, los intervalos de confianza son:

$$\text{-0.65 - 0.79 y -0.65 + 0.79}$$
$$\text{-1.44 y 0.14}$$

Por ello, tenemos el 95% de confianza de que la media de la población caiga entre -1.44 y 0.14 (note que 0 cae en este rango). Sin embargo, así como hay un 5% de oportunidad de cometer un error tipo I, hay también una probabilidad del 5% de que los límites de confianza que se han obtenido no incluyan la media de la población. En efecto, se espera que en promedio, cuando α = 0.05, una de cada 20 muestras resulte en un intervalo de confianza que no incluye Φ.

Los intervalos de confianza son una estimación de la precisión de la estimación que se está estudiando: a menores tamaños de intervalos, mayor la precisión de la estimación. Los IC pueden reducirse en tamaño ya sea reduciendo la varianza de los datos (y por lo tanto el error estándar) o incrementando el tamaño de la muestra (reduciendo tanto el error estándar como el valor crítico de t). Dado que normalmente no se puede reducir la varianza de una población, el método usual para hacer los intervalos de confianza menores es incrementar el tamaño de las muestras.

Los límites de confianza pueden hacerse más amplios (así, más probables de que contengan la media de la población) disminuyendo el valor de α. Si se usa P = 0.001 en vez de 0.05, los intervalos de confianza tendrán un oportunidad de 99.9% de contener la media de la población. Para ν = 11, $t_{0.001(2),11}$ es 4.437. Los límites de confianza para el ejemplo anterior serían:

$$\text{-2.247 y 0.947}$$

Se puede calcular también los IC para el problema de la comparación de medias de dos muestras (Capítulo V). En aquel caso, no se estaba interesado en los valores de las medias de las dos poblaciones, sino que se quería saber el valor que expresa la diferencia entre las dos medias. En ese caso el IC describirá el intervalo que con mayor probabilidad contendrá la diferencia entre las medias de las dos poblaciones.

$$t = \frac{\overline{X}_1 - \overline{X}_2}{S_{\overline{X}_1 - \overline{X}_2}} \rightarrow IC = S_{\overline{X}_1 - \overline{X}_2} \times t_{\alpha(2),\nu} \qquad (6.3)$$

Del Capítulo V:

$$\overline{X}_1 = 34.46$$

$$\overline{X}_2 = 57.25$$

$$S_{\overline{X}_1 - \overline{X}_2} = 4.143$$

$$\overline{X}_1 - \overline{X}_2 = -22.80$$

$$t_{0.05(2),17} = 2.110$$

$$IC = 4.14 \times 2.110 = \mathbf{0.74}$$

Los límites de confianza para la diferencia $\overline{X}_1$ - $\overline{X}_2$ son: -31.54 y -14.05

Nótese que 0 (la no diferencia entre las medias) no cae entre los límites de confianza en este ejemplo.

Suposiciones y Pruebas Inferenciales

En este momento se tiene una herramienta estadística muy poderosa a disposición. Sin embargo, como con cualquier otra herramienta, a menos que se aprenda cómo y cuándo (y cuándo no) usarla, y sus limitaciones, se corre el riesgo de sacar conclusiones inapropiadas de los datos.

Todas las pruebas inferenciales están basadas en un número de suposiciones acerca de los datos. Estas son necesarias para poder hacer comentarios precisos sobre probabilidades. Sin ellas no se puede, por ejemplo, explícitamente definir α, o la probabilidad de un error tipo I. Sin establecer α, no se pueden hacer inferencias estadísticas.

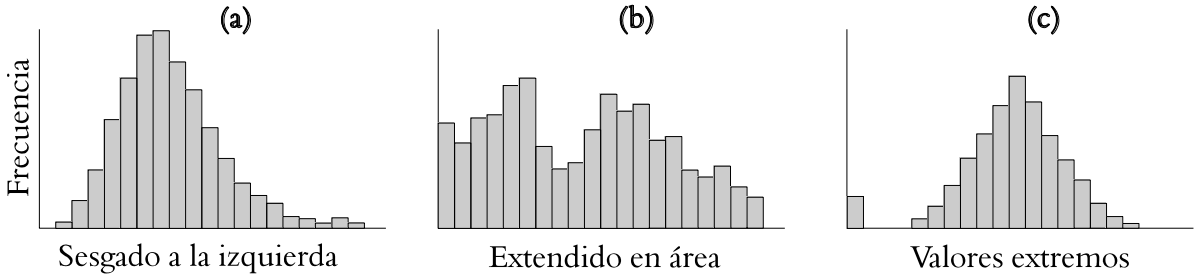
Todas las pruebas inferenciales requieren muestreo al azar de los datos (véase el Capítulo V). Aunque no siempre es necesario que todos los aspectos de un diseño experimental necesiten (o deban) ser aleatorios, debe siempre haber al menos un elemento de aleatoriedad en el diseño.

Muchas pruebas asumen que los datos en cuestión están basados en alguna distribución teórica - generalmente la distribución normal - (algunos, que no cubriremos en este libro, están basados en otras distribuciones). Colectivamente estas pruebas se conocen como **pruebas paramétricas**. La otra gran clase de pruebas inferenciales que se discutirán son las conocidas como pruebas **no paramétricas**.

En adición a la suposición de 'normalidad', muchas pruebas paramétricas (ej. prueba t, ANOVA) también asumen que las varianzas de los diferentes grupos son iguales - la suposición de la homogeneidad de la varianza - y que no hay ninguna correlación entre las magnitudes de la media y la varianza (correlación y regresión). Si cualquiera de esas suposiciones es violada por los datos, es posible obtener resultados 'significativos' que son el producto de las suposiciones violadas. Dado este peligro, la regla No. 1 del análisis de datos es: **mirar los datos y comprobar las suposiciones antes de realizar más pruebas**.

La mejor manera para chequear si los datos no siguen una distribución normal es construyendo un histograma de frecuencias de los datos. Si la forma de la distribución es inclinada (sesgada) a un lado o al otro, muy angosta o plana, demasiado amplia, con valores extremos (véase el Gráfico 6.1) entonces es probable que los datos no estén distribuidos normalmente (véase más adelante qué se puede hacer en ese caso).

Gráfico 6.1



Otra forma de mirar los datos es construir un histograma de frecuencias acumulativas sobrepuesto a los datos (véase el Gráfico 2.5). De nuevo, si se ven diferencias obvias con la curva normal, entonces los datos no están distribuidos normalmente. Hay al menos dos pruebas que pueden ser usadas para probar **diferencias significativas** de normalidad: la prueba del Chi-cuadrado, y la prueba de '**Goodness of Fit' de Kolmogorov-Smirnov**.

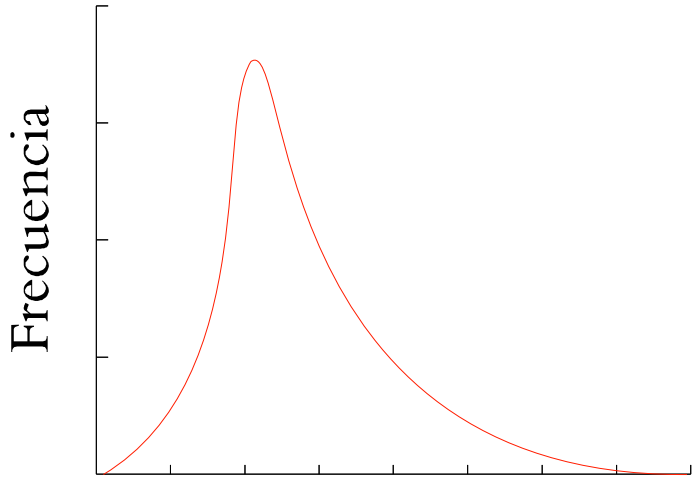
También hay casos en los que datos no están distribuidos normalmente por razones teóricas. Los datos basados en porcentajes y proporciones tienen una distribución **binomial** en vez de una normal. Las distribuciones binomiales difieren en gran medida de las normales para porcentajes pequeños (0-30%) y para grandes (70-100%). Para analizar estos datos normalmente se usa la siguiente transformación:

(6.4)

dónde: sen^{-1} es la función trigonométrica arcoseno, y **p** es un valor de proporción (0 - 1.0).

Muchas distribuciones son atenuadas a la izquierda porque los valores menores que 0 no son posibles (una distribución normal **verdadera** va desde -∞ hasta +∞). Cualquier medida del cuerpo podría caer en esta categoría. Tales distribuciones tienden a ser inclinadas a la (véase el Gráfico 6.2). Muy a menudo tales asimetrías pueden ser corregidas usando una transformación **LOG(X)** o **LOG (X+1)**.

Gráfico 6.2



En algunos casos no será posible corregir los datos por medio de transformaciones. Esto sucede cuando hay valores extremos en uno o ambos lados de la distribución y cuando los datos están ampliamente dispersos sin punto pico central. Entonces probablemente es necesario usar pruebas no paramétricas.

La otra preocupación cuando se hacen pruebas paramétricas es la suposición de homogeneidad de varianzas; esto es que la varianza entre grupos (por ej., los dos grupos en la prueba t para dos muestras) son iguales. Afortunadamente la mayoría de las pruebas paramétricas son robustas frente a pequeñas diferencias en las varianzas sobre todo cuando el tamaño de las muestras son iguales. En el caso de grandes diferencias uno debe considerar el transformar los datos (LOG(X) o LOG (X+1)). Si después de la transformación la varianzas todavía no son iguales entonces se debe considerar una prueba no paramétrica.

Pruebas Paramétricas versas Pruebas No Paramétricas

Las suposiciones de normalidad y homogeneidad de varianzas son suposiciones de pruebas **paramétricas** porque estas pruebas las requieren para ser válidas. Las pruebas no paramétricas o de distribución libre no se basan en suposiciones basadas en alguna distribución. Para la mayoría de las pruebas paramétricas comúnmente usadas existe una prueba no paramétrica que es equivalente.

Si las pruebas no paramétricas son más fáciles de usar (no hay que preocuparse de todas esas suposiciones sobre normalidad) ¿Por qué no se usan todo el tiempo? Una de las razones es que las pruebas paramétricas son siempre más poderosas que su prueba no paramétrica equivalente. Por más poderosas se quiere decir que con las pruebas paramétricas se cometen menos errores del tipo II. En otras palabras, las pruebas paramétricas nos revelan las diferencias reales más fácilmente que las no paramétricas. Este ‘poder’ adicional de las pruebas paramétricas proviene de las suposiciones sobre normalidad y varianza.

Mucha gente usa solamente pruebas paramétricas (por desconocimiento) o no paramétricas (para ser conservador y no cometer errores por las suposiciones). La mejor estrategia es probablemente, usar pruebas paramétricas cuando los datos lo permiten y no paramétricos en caso contrario.

La mayoría de las pruebas no paramétricas están basadas en los rangos de los datos más que en los datos mismos. El primer paso en estas pruebas es el organizar los datos (ya sea de forma ascendente o descendente) y entonces asignar un rango del menor al mayor (más sobre esto después). Debido a que se usan los rangos de los datos en lugar de los datos, se pierde información (parecido a la mediana que utiliza mucha menos información que la media). Como resultado, las pruebas no paramétricas son menos eficientes que las paramétricas. Por otro lado, sin embargo, el utilizar rangos hace a las pruebas no paramétricas muy frente los valores extremos e las diferencias en varianzas. Por ello, es mucho más fácil usar pruebas no paramétricas con datos de la escala ordinal.

La prueba U de Mann-Whitney

Para ilustrar el uso de las pruebas no paramétricas, se usará la prueba no paramétrica equivalente a la prueba t para dos muestras: la **prueba U de Mann-Whitney**.

Los datos de Tabla 6.1 corresponden a diámetros de árboles tomados a una altura de 1 m. en dos diferentes lugares de Santa Cruz (véase Tabla 6.1). La pregunta es: "¿En promedio, son los diámetros de las dos áreas iguales (área A y área B)?"

Se pueden asignar rangos tanto del más pequeño al más grande, como viceversa. En el siguiente ejemplo los rangos han sido asignados del más pequeño al más grande. Al valor más pequeño (8) se le ha dado el valor de 1. Al valor que le sigue el de 2 y así sucesivamente. El valor más grande tiene el valor de rango de 13, que es también el número total de observaciones.

El estadístico U de Mann-Whitney se calcula:

$$
\begin{aligned}
U &= n_1 \times n_2 + \frac{n_1 \times (n_1 + 1)}{2} - R_1 \qquad (6.6) \\
&= 8 \times 5 + \frac{8 \times (8 + 1)}{2} - 67 \\
&= 40 + 36 - 67 \\
&= \mathbf{9}
\end{aligned}
$$

Un estadístico también se debe calcular:

$$U' = n_1 \times n_2 - U \qquad (6.7)$$
$$= 8 \times 5 - 9$$
$$= 40 - 9$$
$$= \mathbf{3}1$$

Tabla 6.1

Diámetros de los árboles (en orden ascendente)			Diámetros Ordenados	
A	B		A	B
10	8		3	1
15	9		5	2
22	13		6	4
30	25		8	7
31	34		9	10
40			11	
64			12	
85			13	
$n_1 = 8$	$n_2 = 5$	Σ Rangos:	$R_1 = 67$	$R_2 = 24$

Al igual que con las pruebas t, se necesita mirar el valor crítico de la prueba. Los valores críticos se encuentran buscando n_1 y n_2 (donde n_1 es ahora el menor de los dos tamaños de muestra y n_2 es el mayor de los dos).

Se debe tener cuidado con la Tabla de valores críticos usada. Hay al menos dos formas de calcular la prueba U de Mann-Whitney y cada una utiliza su propia Tabla de valores críticos.

El valor crítico $U_{0.05(2),5,8}$ en el ejemplo que se está mostrando es 34.

Si U o U' es mayor o igual a se rechaza H_o y se acepta H_a. Si ambos U y U' son menores que se falla en rechazar H_o.
En el ejemplo, ambos, U (9) y U'(31), son menores que $U_{0.05(2),5,8}$ (34) y por lo tanto no se rechaza H_o.

El cálculo de U y U' puede también lograrse intercambiando R_1 y R_2 como se muestra a continuación:

$$U = n_2 \times n_1 + \frac{n_2 \times (n_2 + 1)}{2} - R_2 \qquad (6.9)$$

$$U' = n_2 \times n_1 - U \qquad (6.10)$$

La prueba U de Mann-Whitney es la más poderosa de las pruebas no paramétricas de dos muestras. Cuando se puede aplicar esta prueba o la t, la prueba U tiene solamente al rededor del 95% del poder de t.

Las técnicas descritas arriba para calcular la prueba U funcionan bien mientras no hayan rangos **empates** en los datos. Cuando estos ocurren el proceso de ordenamiento se hace un poco más complicado. Consideremos el siguiente ejemplo:

Tabla 6.2

A	B	Rango A	Rango B
10	8	3	1
11	9	4.5	2
13	11	7	4.5
13	15	7	10
13	16	7	11
14	20	9	13
20	20	13	13
30	28	16	15
$n_1 = 8$	$n_2 = 8$	$R_1 = 66.5$	$R_2 = 69.5$

Como antes, a los tres valores menores (8, 9, 10) se les da rangos de 1, 2 y 3 respectivamente. Sin embargo, hay dos observaciones con el próximo valor en tamaño 11. Para asignar rangos a valores empatados se necesita calcular el **rango promedio** para las observaciones. El primer paso es sumar los rangos de las observaciones como si fueran diferentes. En el caso de los 11, los rangos serían 4 y 5, que suman 9. Ahora se divide esta suma por el número de observaciones de valor 11, en este caso 2. Y se obtiene que el rango promedio para los 11's es 4.5. El próximo valor en los datos es 13 y hay tres observaciones con este valor. La suma de los rangos sería 6 + 7 + 8 = 21, que dividido por 3 es 7. Si se sigue el mismo proceso para los 20 el rango promedio es 13).

Una vez se completa el ordenamiento, el resto de la prueba se lleva a cabo como antes:

$$U = 8 \times 8 + \frac{8 \times (8+1)}{2} - 66.5$$

$$= 64 + 36 - 66.5$$

$$= \mathbf{33.5}$$

$$U' = 8 \times 8 - 33.5$$

$$= \mathbf{30.5}$$

Valor Crítico: $U_{0.05(2),8,8} = 51$

Como 33.5 y 30.5 son ambos menores que el valor crítico 51, no se rechaza la hipótesis nula de que los dos grupos son iguales.