5A. ESTADÍSTICA INFERENCIAL.

SELECCIÓN, ORDENACIÓN Y SEGMENTACIÓN DE DATOS.

ESTADÍSTICA INFERENCIAL

Ofrece información de la población origen, a partir de los datos obtenidos en la muestra del estudio. Trata de la generalización hacia las poblaciones de los resultados obtenidos en las muestras. Sus dos pilares básicos son la *Estimación de parámetros* y las *Pruebas de significación y Contraste de hipótesis*.

Estimación de parámetros

La estimación más intuitiva de un parámetro de la población, p. ej. la media de *Imc*, es la media concreta de *Imc* de la muestra, 29,3 m/kg^2 en este caso (Tabla 4A.2). Esta es una estimación puntual. Al realizar el estudio con otras muestras distintas de igual tamaño, la estimación puntual de la media sería un valor distinto en cada muestra. Además, debido a la variabilidad aleatoria, es muy poco probable que la estimación puntual de un parámetro acierte con el verdadero valor en la población. Por estos motivos es más adecuado realizar la estimación de parámetros por intervalo: se ofrece un rango de valores entre los cuales hay una probabilidad alta de encontrar el verdadero valor del parámetro en la población. Habitualmente se acepta esta probabilidad en el 95%, por lo que el *Intervalo de confianza del 95%* (IC95%) es la estimación de parámetros más utilizada.

En la muestra de 112 casos, la media de Imc es 29,3 m/kg^2 con una DE de 4,4 m/kg^2 (Tabla 4A.2). La población origen de la muestra tiene una media M_P desconocida, que es la queremos estimar, y una desviación estándar DE_P también desconocida. Si obtenemos infinitas muestras de 112 casos de la misma población y repitiéramos el estudio, obtendríamos infinitas medias de Imc, una de cada muestra. No todas las medias tendrán la misma probabilidad de salir. La más probable sería la verdadera media de la población M_P. Los valores próximos a ambos lados de esta media poblacional M_P, tendrán menos probabilidad de salir, y a medida que los valores se alejan de esa media poblacional, disminuye su probabilidad de salir. Es decir la distribución de probabilidad de todas las posibles muestras, llamada *distribución muestral*, sigue una ley normal de media M_P y desviación estándar $DE_P/\sqrt{n}$, dependiente directamente de la DE_P e inversamente del tamaño muestral, n = 112 en este caso. A esta estimación de la desviación estándar de la distribución muestral ($DE_P/\sqrt{n} = 4{,}4/\sqrt{112} = 0{,}4$) es a lo que se llama *Error Estándar* (EE). Según las leyes de la ley normal, en el intervalo $M_P \pm 2*DE_P/\sqrt{n}$ se encuentran el 95% central de las medias muestrales. Pero sólo se dispone de una muestra y de un resultado, con el que se puede construir un intervalo ($29{,}3 \pm 2*4{,}4/\sqrt{112} = 29{,}3 \pm 2 *0{,}4 \rightarrow 28{,}5$ a $30{,}1$) que tendrá una confianza del 95% de contener el verdadero valor de la media poblacional. Es el Intervalo de confianza del 95%: la media de Imc de la población origen de la muestra esta comprendida entre 28,5 y 30,1 m/kg^2, con una confianza del 95%.

Es importante tener claros los conceptos de Desviación Estándar, Error Estándar, intervalo de Normalidad del 95% e Intervalo de confianza del 95%:

– **Desviación Estándar** (DE). Es una medida de la variabilidad de los valores de Imc de los sujetos de la muestra (DE = 4,4 m/kg^2). Con respecto a su comunicación en las publicaciones, aunque carece de interpretación práctica, puede ser útil al lector para el cálculo del tamaño de muestra o para un metanálisis.

– **Intervalo de normalidad** del 95% (IN95%). Si la variable sigue una ley normal, el siguiente intervalo contendrá los valores de Imc del 95% central de los sujetos de la muestra: $M \pm 2*DE = 29{,}3 \pm 2*4{,}4 \rightarrow$ IN95%: 20,5 a 38,1 kg/m^2.

– **Error Estándar** (EE). Es una medida de la variabilidad de los valores de la media de Imc de la distribución muestral de medias (EE = 0.4 m/kg^2). Es una medida de la variabilidad de los errores de muestreo de las medias, que valora la precisión con que la media de la muestra estima la media de la población.

– **Intervalo de Confianza** del 95% (IC95%). Es el intervalo que tiene un 95% de probabilidades de éxito de contener la verdadera media de Imc de la población. Cuando la muestra es grande el IC95% de la media se obtiene: $M \pm 2*EE = 29{,}3 \pm 2*0{,}4 \rightarrow$ IC95%: 28,5 a 30,1 m/kg^2. Es más reducido que el IN95%. A la hora de comunicarlo en una publicación, no se recomienda indicar el EE, es mejor indicar el error de muestreo en forma de IC95%, que lleva implícito el EE.

La DE y el EE son medidas matemáticas de variabilidad de difícil interpretación práctica. Es más fácil interpretar estas variabilidades expresándolas en forma de rangos de valores: el IN 95% indica la variabilidad de los valores de los sujetos (DE); el IC indica la variabilidad de los errores de muestreo de la media (EE) acotando la zona donde se sitúa la media de la población.

En el videotutorial **3A2Explorar1.avi** se puede observar como la tabla *Descriptivos* del cuadro *Explorar* ofrece como medidas de estadística inferencial, el EE y el IC95% de la media y los EE de los índices de Asimetría y Curtosis.

Pruebas de significación y contraste de hipótesis

Para ver si la presencia o ausencia de *Obesidad* se asocia a diferentes niveles de *Glucosa*, se analizan los datos obtenidos en la muestra. En los obesos la media de glucemia es $M_O = 175$ mg/dl y en los no obesos es $M_N = 143$ mg/dl (Tabla 10A.1). La diferencia de medias es $M_O - M_N = 175 - 143 = 32$ mg/dl (Tabla 10A.2). Si se realizara el mismo estudio con varias muestras distintas de 112 individuos extraídos de la misma población, el resultado sería diferente en cada muestra debido al error aleatorio de muestreo. Por tanto cabe preguntarse si la diferencia encontrada en la muestra de 32 mg/dl ¿es una estimación más o menos válida de la verdadera diferencia existente en la población? o ¿es una diferencia explicable por el azar en una muestra extraída de una población en la que no existen diferencias? Solo hay dos soluciones excluyentes a esta cuestión: o las medias de la población son iguales con lo que su diferencia es nula (Hipótesis nula H_0: $M_O - M_N = 0$) o las medias de la población son diferentes con lo que su diferencia es distinta de cero (Hipótesis alternativa H_1: $M_O - M_N \neq 0$).

Para dar respuesta a este dilema la estadística aprovecha la única información de que dispone: la distribución de probabilidad de las diferencias de medias de todas las posibles muestras en el caso de que no haya diferencias (hipótesis nula). La distribución de probabilidad en el caso de diferencias de medias la ofrece la ley t de Student. Hay una ley t de Student para cada tamaño de muestra o grados de libertad. La tablas de estadística que se utilizaban antiguamente o los programas estadísticos de ordenador actuales ofrecen el resultado de este ejemplo concreto: el estadístico $t = 2{,}365$, los grados de libertad $n-2$ gl $= 110$ y el nivel de significación $p = 0{,}02$. La probabilidad de obtener una diferencia de medias ≥ 32 mg/dl en muestras de 112 casos extraídas de una población con diferencia de medias nula es de 2% ($p = 0{,}02$). En base a la magnitud de esta probabilidad (grande o pequeña) se toma una decisión. Arbitrariamente se establece el 0,05 como límite entre probabilidad grande y pequeña. Si la probabilidad es grande ($p > 0{,}05$), la diferencia encontrada es explicable por el azar, se acepta la hipótesis nula y por tanto se asume que no hay diferencias estadísticamente significativas en las medias de glucemia en obesos y en no obesos. Si la probabilidad es pequeña ($p \leq 0{,}05$) la diferencia encontrada es poco explicable por el azar, por lo que no se puede aceptar la hipótesis nula y por tanto, por exclusión, se asume como verdadera la hipótesis alternativa: hay diferencias estadísticamente significativas en las medias de glucemia en obesos y en no obesos. Es el caso de este ejemplo.

Este es el fundamento básico de todas las pruebas de significación estadísticas: hay una hipótesis nula, que refleja la ausencia de relación entre las variables, y otra alternativa, en la que hay relación entre las variables. La única hipótesis que se pone a prueba es la hipótesis nula porque es la única de la que se dispone de distribución de probabilidad. Es esta distribución de probabilidad la que cambia de una prueba a otra (p. ej. la Chi-cuadrado en la relación entre dos variables categóricas, la F de Snedecor en la comparación de dos varianzas, etc.), pero el fundamento subyacente en todas las pruebas estadísticas es el mismo.

Las pruebas de significación estadística funcionan en base a la probabilidad de un resultado. Por tanto se pueden dar cuatro situaciones posibles. Dos de ellas conducen a conclusiones correctas y otras dos a conclusiones equivocadas (Tabla 5A.1).

Tabla 5A.1. Pruebas de significación y contraste de hipótesis.

		Prueba estadística	
		No significativa $p > 0{,}05$ Se acepta H_0	Significativa $p \leq 0{,}5$ Se rechaza H_0
Relación real	Hay relación H_1 verdadera	Falso negativo **Error aleatorio tipo II** Riesgo β	Verdadero positivo **Potencia** 1-β
	No hay relación H_0 verdadera	Verdadero negativo **Confianza** 1-α	Falso positivo **Error aleatorio tipo I** Riesgo α

– **Riesgo alfa**. Cuando no hay relación real y la prueba es significativa (falso positivo) se produce un error aleatorio tipo I y su probabilidad de ocurrencia es el riesgo alfa. Alfa es la probabilidad de dar por verdadera una relación que no existe en la realidad. Está cuantificado en 5% al adoptar el límite arbitrario de significación en el 0,05.

– **Confianza**. Cuando no hay relación real y la prueba es no significativa (verdadero negativo) se produce una conclusión correcta llamada nivel de Confianza y su probabilidad de ocurrencia es 1−alfa, habitualmente el 95%, para alfa de 0,05.

– **Riesgo beta**. Cuando hay relación real y la prueba es no significativa (falso negativo) se produce un error aleatorio tipo II y su probabilidad de ocurrencia es el riesgo beta. Beta es la probabilidad de no detectar una relación que en realidad existe. Se suele cuantificar en 0,20 (20%).

– **Potencia**. Cuando hay relación real y la prueba es significativa (verdadero positivo) se produce una conclusión correcta llamada Potencia de la prueba y su probabilidad de ocurrencia de 1−beta, habitualmente el 80%, para beta de 0,20.

La mayoría de la estadística que se encuentra en la bibliografía médica se refiere a la probabilidad de error alfa y se expresa mediante el familiar valor *p* (significación estadística). El valor *p* no es más que la probabilidad de encontrar diferencias iguales o superiores a las que han salido en la muestra, en el caso de que no exista diferencia real en la población. Indica la probabilidad de que las diferencias encontradas en la muestra sean debidas al azar, y si la relación que se ha encontrado entre las variables de la muestra se puede asumir o no en la población origen de la muestra. Cuando se llega a la conclusión de que hay diferencias ($p \leq 0,05$) este valor de *p* cuantifica bien la probabilidad de error que se comete (alfa). Pero cuando la conclusión del estudio es que no hay diferencias ($p > 0,05$), el valor de *p* no es tan importante y es el riesgo beta el que cuantifica la probabilidad de error que se comete. En la literatura médica se ha prestado menos atención al riesgo beta que al riego alfa. Las razones son que es más difícil de calcular y que en general se prefieren los estudios con resultados positivos. Incluso en estudios con resultados negativos se ofrecen argumentos diferentes del azar para explicar los resultados, o bien se hace hincapié en las diferencias encontradas en subgrupos, aunque no sean estadísticamente significativas, cuando lo correcto es indicar cuál es la probabilidad de falso negativo o riesgo beta. El riesgo beta es especialmente alto en el caso de muestras pequeñas (estudios con poca potencia). Este problema no se plantea si previamente al inicio del estudio se ha realizado la predeterminación del tamaño muestral.

Actualmente los valores de *p* se obtienen con programas estadísticos informáticos que dan los valores exactos. Es recomendable ofrecer este valor de *p* exacto (p. ej., 0,03, 0,07, 0,11) para dar una idea más precisa de lo que por azar es más o menos probable y contribuir a la interpretación de lo que es estadísticamente significativo, dada la arbitrariedad del valor 0,05. Sin embargo es recomendable poner límites a estas probabilidades exactas. Así, valores de $p > 1/5$ suelen reseñarse como $> 0,20$ ya que indican una probabilidad inaceptablemente elevada. Del mismo modo, valores de *p* muy reducidos se expresan como $p < 0,001$, indicando que es muy poco probable que la diferencia sea explicada por el azar.

Prueba estadística unilateral y bilateral

La prueba de significación descrita y formulada con las Hipótesis nula (H_0: $M_O - M_N = 0$) y alternativa (H_1: $M_O - M_N \neq 0$) es una prueba *bilateral*. Las diferencias pueden producirse en ambos lados de la igualdad: la obesidad puede aumentar o disminuir la glucemia. La prueba estadística es *unilateral* cuando las hipótesis se formulan así: Hipótesis nula H_0: $M_O - M_N \leq 0$; Hipótesis alternativa H_1: $M_O - M_N > 0$. Se asume que la obesidad sólo puede aumentar la glucemia. Las pruebas unilaterales son más "ventajosas" para el investigador que las bilaterales porque tienen valores de *p* menores, habitualmente la mitad. Pero se debe tener gran conocimiento teórico del tema objeto de estudio y argumentar claramente que se tiene la seguridad de que no es posible el resultado en sentido contrario, que la obesidad disminuye la glucemia en el ejemplo. Esto pocas veces es posible, por lo que por regla general se utilizan pruebas estadísticas bilaterales.

Predeterminación del tamaño muestral

Consiste en calcular cuántos sujetos son necesarios para conseguir que los resultados sean lo más válidos posibles con una probabilidad baja y predeterminada de ser explicados por el azar. El tamaño de muestra de un estudio depende de cinco factores. Para el estudio de comparación de medias de glucemia entre obesidad y no obesidad del ejemplo queremos predeterminar el tamaño de muestra necesario especificando esos cinco factores:

– ***Magnitud del efecto***. Cuanto menor sea la diferencia entre los grupos que quiera detectarse, mayor será el tamaño de muestra requerido. Esta diferencia es el menor efecto con importancia clínica o práctica. Es una estimación realizada por el investigador en base a conocimientos teóricos o estudios previos publicados. En el ejemplo, por opinión de expertos en diabetes o según revisión de la literatura médica sobre el tema, se estima que el 10 mg/dl de diferencia de glucemia entre obesos y no obesos, es la mínima diferencia con importancia clínica en cuanto a desarrollo de complicaciones.

– ***Riesgo alfa***. Cuanto menor sea el riesgo alfa, mayor será el tamaño de muestra. Habitualmente se sitúa en 0,05 (1/20), excepcionalmente en 0,01 (1/100). Se da gran valor a estar seguros de que un efecto está presente cuando realmente lo está.

– ***Riesgo beta***. En general se admite un riesgo beta < 20% (excepcionalmente 10%) y una potencia > 80% (excepcionalmente 90%). Es la probabilidad de pasar por alto diferencias reales. Suele ser mayor que el alfa, por lo que se da mas valor a decir que un efecto existe cuando realmente está presente (riesgo alfa menor) y menos valor a decir que no existe cuando realmente no está presente (riesgo beta mayor).

– ***Características de los datos***. Para datos binarios depende de las proporciones relativas: con 50%−50% se requieren menos casos que a medida que esas proporciones se alejan. Con datos cuantitativos dependen directamente de la dispersión de los datos: a mayor dispersión mayor muestra y viceversa. En el ejemplo depende de la varianza o desviación estándar de la glucemia en ambos grupos. Se puede obtener de estudios previos similares o de un estudio piloto con pocos pacientes. Si los grupos son de tamaño distinto y las varianzas diferentes, se obtiene la media ponderada por los grados de libertad de las varianzas o desviación estándar de ambos grupos. En el ejemplo, asumimos que ambos grupos son iguales y con la misma varianza o desviación estándar cuyo valor estimamos a partir de la muestra global de 112 casos del ejemplo, cuyo resultado es varianza de 5117 $(mg/dl)^2$ y desviación estándar de 71,5 mg/dl (Tabla 4A.2).

– ***Prueba estadística unilateral o bilateral***. Las pruebas bilaterales necesitan mayor muestra que las unilaterales.

La predeterminación del tamaño muestral debe hacerse antes de iniciar el trabajo. En muchas ocasiones son necesarios demasiados sujetos para los medios de los que se dispone para realizar la investigación, en cuyo caso se debe abandonar el proyecto inicial o adaptarlo a la disponibilidad de medios.

El tamaño de muestra se obtiene con fórmulas matemáticas, más o menos complejas, o directamente de programas de ordenador. En nuestro caso, para comparar, con prueba bilateral, las medias de glucemia entre dos grupos (obesidad y no obesidad) con una magnitud del efecto estimada en una diferencia de glucemia mínima de 10 mg/dl, asumiendo un error alfa del 5% y un error beta del 20% y con una varianza estimada de 5117 $(mg/dl)^2$ (desviación estándar de 71,5 mg/dl), necesito un tamaño de muestra de 804 sujetos por grupo, 1608 en total (Videotutorial 5A1Muestra.avi). Es una cifra que podría hacer reconsiderar el estudio analizando los medios de que se dispone.

Videotutorial 5A1Muestra.avi
Se muestra como se obtiene, con el programa para análisis epidemiológico de datos Epidat, el tamaño de muestra para comparar las medias de glucosa en dos grupos (obesos y no obesos) de igual tamaño asumiendo varianzas iguales, un riesgo alfa del 5% (nivel de confianza del 95%), riesgo beta del 20% (potencia del 80%) y una desviación estándar común estimada en 71,5 mg/dl.

Pruebas estadísticas para diseños de grupos independientes

La respuesta a la pregunta de investigación de un estudio se alcanza analizando la relación entre las variables exposición y respuesta observadas en una muestra. Se dice que las dos variables están relacionadas si la variable respuesta se va modificando a medida que la variable exposición cambia sus valores. Si la respuesta no se modifica cuando cambia la exposición, se dice que las variables no están relacionadas. El análisis de esta relación se realiza mediante la utilización de las pruebas estadísticas para estimar la probabilidad de riesgo alfa. El tipo concreto de prueba estadística a utilizar depende de la escala de medida de las variables exposición y respuesta y de que los datos cumplan algunas asunciones que se detallarán en el desarrollo de cada prueba.

El archivo *Factores.sav* contiene, entre otros, los datos sobre Obesidad y Diabetes de 112 sujetos. Se pretende analizar la relación entre la exposición Obesidad y la respuesta Diabetes. Obesidad está recogida como binaria o dicotómica (*Obesidad*), categórica con más de dos categorías o politómica (*ObesidadOr*) y cuantitativa (*Imc*). Diabetes está como binaria (*Diabetes*) y como cuantitativa (*Glucosa*) y podría estar como categórica con más de dos categorías (*DiabetesOr*). La variable *DiabetesOr* se podría llamar "Nivel de Glucosa" y tendría 3 categorías codificada de la siguiente forma: 0 = No diabetes (Glucosa < 126 mg/dl), 1 = Diabetes leve (Glucosa ≥ 126 y < 200 mg/dl) y 2 = Diabetes severa (Glucosa ≥ 200 mg/dl). También se podría haber recogido la variable *Tiempo a Diabetes* como el tiempo transcurrido hasta el diagnóstico de diabetes o el tiempo de seguimiento hasta la última observación en caso de no haber sido diagnosticado de diabetes.

Para valorar la relación entre la exposición Obesidad y la respuesta Diabetes, el análisis más adecuado es utilizar las variables cuantitativas *Imc* y *Glucosa*. Con fines didácticos también utilizaremos las variables categóricas *Obesidad*, *ObesidadOr*, *Diabetes* y *DiabetesOr* con el objetivo de poder desarrollar las pruebas estadísticas más importantes utilizadas en los diseños de grupos independientes, según se muestra en la Tabla 5A.2. En rojo se muestran las pruebas que se verán en el presente curso. En azul las prueban que se ven en el curso "Supervivencia y Modelos de Regresión Lineal, Logística y de Cox", más avanzado y continuación de este.

Variable Respuesta Binaria (*Diabetes*)

- **Variable Exposición Binaria** (*Obesidad*). Es la *comparación de 2 proporciones*: proporción de diabéticos en obesos frente a proporción de diabéticos en no obesos. Se realiza con la significación estadística de la *prueba de chi-cuadrado (χ^2)* entre dos variables binarias. Si no se cumplen sus condiciones de aplicación se utiliza la *prueba exacta de Fisher*. La valoración práctica y clínica se realiza mediante las medidas de asociación: Riesgo Relativo (RR) y Razón de Odds (RO) u Odds Ratio (OR).
- **Variable Exposición Categórica con más de dos categorías** (*ObesidadOr*). Es la *comparación de varias proporciones*: proporción de diabéticos en obesos, proporción de diabéticos en sobrepeso y proporción de diabéticos en Imc normal. Se realiza con la *prueba de chi-cuadrado (χ^2)* entre una variable binaria y una categórica con más de dos categorías. Es una generalización de la prueba anterior.

- **Variable Exposición Cuantitativa** (*Imc*). La comparación se realiza con la *Regresión Logística Binaria*, y estaríamos ante un problema de clasificación consistente en estimar un modelo que clasifique a los pacientes en diabéticos o no diabéticos a partir de su índice de masa corporal. La magnitud del efecto se valora con la Razón de Odds (RO) de la variable respuesta (Diabetes), por cada unidad de aumento de la variable exposición (por cada kg/m^2 de aumento de Imc). La Regresión Logística Binaria es el modelo general cuando la variable respuesta es binaria, independientemente de como sean las variables exposición (binarias, categóricas con más de dos categorías o cuantitativas). En el caso de variables exposición categóricas los resultados son similares a los obtenidos con las pruebas de chi-cuadrado y medidas de riesgo.

Variable Respuesta Categórica con más de dos Categorías (*DiabetesOr*)

- **Variable Exposición Binaria** (*Obesidad*). Es la *comparación de varias proporciones* que se realiza con la significación estadística de la *Prueba de chi-cuadrado (χ^2)* entre una variable respuesta categórica con más de dos categorías (*DiabetesOr*) y otra exposición binaria (*Obesidad*).
- **Variable Exposición Categórica con más de dos categorías** (*ObesidadOr*). Es la *comparación de varias proporciones* que se realiza con la *Prueba de χ^2* entre dos variables categóricas con más de dos categorías, en este caso la variable respuesta categórica *DiabetesOr* y la variable exposición categórica *ObesidadOr*.
- **Variable Exposición Cuantitativa** (*Imc*). La comparación se realiza con la *Regresión Logística Multinominal* (si la respuesta es nominal) o con la *Regresión Logística Ordinal* (si la respuesta es ordinal como en este caso) y estaríamos ante un problema de clasificación consistente en estimar un modelo que clasifique a los pacientes en una de las tres categorías de *DiabetesOr* a partir de su índice de masa corporal. La Regresión Logística Multinominal u Ordinal es el modelo general cuando la variable respuesta es categórica con más de dos categorías, independientemente de como sean las variables exposición (binarias, categóricas con más de dos categorías o cuantitativas).

Tabla 5A.2. Principales pruebas estadísticas para diseños de grupos independientes.

VARIABLE RESPUESTA	VARIABLE EXPOSICIÓN BINARIA (Obesidad)	 CATEGÓRICA (ObesidadOr)	 CUANTITATIVA (Imc)	MODELO GENERAL
BINARIA (Diabetes)	**Prueba de χ^2** Prueba de Fisher (comparación de dos proporciones)		Regresión Logística Binaria Simple	**Regresión Logística** Binaria Múltiple
CATEGÓRICA (DiabetesOr)	Prueba de χ^2 (comparación de varias proporciones)		Regresión Logística Multinominal u Ordinal Simple	Regresión Logística Multinominal u Ordinal Múltiple
CUANTITATIVA CONTINUA (Glucosa)	**t de Student-Fisher** U de Mann-Whitney (comparación de dos medias)	**Análisis de la varianza** Kruskal-Wallis (comparación de varias medias)	Correlación de Pearson y de Spearman Regresión Lineal Simple (comparación de 2 variables cuantitativas)	**Regresión lineal** Múltiple
SUPERVIVENCIA (Diabetes + Tiempo Diabetes)	**Método Kaplan-Meier: Prueba de Log Rank (Mantel-Cox)** (comparación de curvas de supervivencia)		Regresión de Cox de riesgos proporcionales Simple	**Regresión de Cox** de riesgo proporcional Múltiple

Variable Respuesta Cuantitativa (*Glucosa*)

- **Variable Exposición Binaria** (*Obesidad*). Es la *comparación de dos medias*: la media de glucosa en obesos y la media de glucosa en no obesos. Se realiza con la *prueba t de Student-Fisher*. En caso de no cumplirse sus condiciones de aplicación se utiliza la prueba no paramétrica *U de Mann-Whitney*. La valoración de la magnitud del efecto o importancia práctica se realiza a través del IC95% de la diferencia entre las medias.
- **Variable Exposición Categórica con más de dos categorías** (*ObesidadOr*). Es la *comparación de varias medias*, las medias de glucosa correspondientes a las tres categorías de *ObesidadOr*: mc Normal, Sobrepeso y Obesos. Se realiza con el *Análisis de la Varianza* (ANOVA). En una generalización de la t de Student para la comparación de varias medias. En caso de no cumplirse sus condiciones de aplicación se utiliza la prueba no paramétrica *H de Kruskal-Wallis*.
- **Variable Exposición Cuantitativa** (*Imc*). Es la *comparación de dos variables cuantitativas*, que se realiza con la ***Regresión Lineal** Simple***.** La interpretación práctica se valora con el *coeficiente de regresión β* que representa la pendiente de la recta de regresión ajustada, e indica la magnitud de cambio de la variable respuesta por cada unidad de aumento de la variable exposición. En este caso representa el cambio (aumento o disminución) de glucosa por cada unidad (kg/m^2) de aumento de Imc. La regresión lineal simple es el modelo general cuando la variable respuesta es cuantitativa: también se puede emplear con variables exposición binarias, dando resultados idénticos a la prueba t de Student y con variables exposición categóricas con más de dos categorías, dando resultados idénticos al análisis de la varianza. Cuando se emplean dos o más variables exposición que pueden ser binarias, categóricas o cuantitativas estamos ante la generalización de esta técnica

denominada *Regresión Lineal Múltiple*. La ***Correlación*** estudia la relación lineal entre dos variables cuantitativas que juegan un papel simétrico (no hay variable exposición y respuesta), utilizando el *coeficiente de correlación r de Pearson* o el *coeficiente de correlación rho de Spearman* no paramétrico.

Variable Respuesta de Supervivencia o Tiempo a un evento (*Diabetes* + *Tiempo a Diabetes*)

- **Variable Exposición Binaria** (*Obesidad*). Es la *comparación de dos curvas de supervivencia*: curva de supervivencia de los obesos frente a curva de supervivencia de los no obesos. Se realiza con la *prueba de log rank* del método de Kaplan-Meier del análisis de supervivencia.
- **Variable Exposición Categórica con más de dos categorías** (*ObesidadOr*). Es la *comparación de varias curvas de supervivencia*: curvas de supervivencia de Imc normal, de Sobrepeso y de Obesidad. Se realiza con la *prueba de log rank* del método de Kaplan-Meier del análisis de supervivencia.
- **Variable Exposición Cuantitativa** (*Imc*). La valoración del efecto de una variable cuantitativa sobre una variable de supervivencia se realiza mediante la *Regresión de riesgos proporcionales de Cox simple*. La magnitud del efecto se valora con la Razón de tasas de incidencia o *Hazard Ratio* (HR) de la variable respuesta (Diabetes) por cada unidad de aumento de la variable exposición (por cada kg/m^2 de aumento de Imc). La regresión de Cox es el modelo general cuando la variable respuesta es de supervivencia: también se puede emplear con variables exposición binarias y categóricas con más de dos categorías, dando resultados similares, no idénticos, a las pruebas de log rank del método de Kaplan-Meier. Cuando se emplean dos o más variables exposición que pueden ser binarias, categóricas con más de dos categorías adecuadamente codificadas o cuantitativas estamos ante la generalización de esta técnica denominada *Regresión de Cox Múltiple*.

El uso de **modelos generales de regresión lineal, logística y de Cox**, es fundamental para valorar simultáneamente varias variables exposición sobre la respuesta. En investigación no experimental permite valorar **variables de confusión** e introducir **variables de interacción**, también llamadas modificadoras del efecto. El análisis multivariante con modelos de regresión y el manejo de la confusión y la interacción se tratan en profundidad en el curso “Supervivencia y Modelos de Regresión Lineal, Logística y de Cox”.

Tipos de pruebas estadísticas para comprobar hipótesis

- **Pruebas de Bondad de ajuste**. Verifican una hipótesis sobre la forma de distribución de la población. Analizan la bondad del ajuste de un conjunto de datos a distribuciones teóricas. Tienen en cuenta si los valores observados difieren significativamente de los datos esperados en el supuesto de que la variable siga una determinada distribución estadística. Ejemplos: las Pruebas de Normalidad de Kolmogorov-Smirnov y de Shapiro-Wilks, para comprobar si el Imc sigue una ley normal. Si la $p \leq 0{,}5$ los datos difieren significativamente de una distribución normal teórica (no cumplen la normalidad). Si la $p > 0{,}5$ los datos no difieren significativamente de una distribución normal teórica (cumplen la normalidad).

- **Pruebas de Conformidad**. Verifican una hipótesis sobre un determinado valor de un parámetro de la población. Ejemplo: la prueba de comparación de una media observada a una teórica (comprobar si la muestra procede de una población con Imc =25 kg/m^2) o la prueba de comparación de una proporción observada a una teórica (ver si la muestra procede de una población con igual proporción de hombres y mujeres).

- **Pruebas de Independencia**. Verificar la asociación entre dos variables que juegan un papel simétrico, la hipótesis no contempla una variable independiente y otra dependiente. Son el objetivo principal en los estudios de asociación, caso frecuente en estudios transversales. Ejemplo: Coeficiente de Correlación entre el Peso y la Talla o prueba de Chi-cuadrado entre Sexo y Hospital.

- **Pruebas de Homogeneidad**. Verificar la relación de igualdad o desigualdad entre una variable exposición y otra respuesta, en la que las variables juegan un papel asimétrico. Son el objetivo principal de los estudio de comparación. Ejemplo: Prueba t de Student de comparación de medias de Glucosa en obesos y no obesos, que sería un estudio comparativo de la respuesta observada en los dos grupos.

- **Pruebas de Tendencia**. Verificar la ausencia o presencia de tendencia creciente o decreciente en un estudio comparativo cuando la variable exposición tiene categorías ordenadas. Ejemplo: Prueba de tendencia lineal del ANOVA entre las medias de Glucosa en las categorías ordenadas de la variable ObesidadOr.

SELECCIÓN, ORDENACIÓN Y SEGMENTACIÓN DE DATOS

Selección de casos

Con frecuencia se necesita realizar análisis estadísticos en una fracción de los casos del archivo de datos. Es lo que se conoce como seleccionar casos o filtrar el archivo. El cuadro *Seleccionar casos* permite seleccionar los casos según una determinada condición lógica, o de forma aleatoria, o según un rango de orden de los casos o del tiempo o según otra variable utilizada como filtro. Los casos no seleccionados pueden ser filtrados temporalmente o eliminados de forma permanente y guardarse en un archivo con otro nombre.

Si queremos hacer una tabla de Frecuencias de las categorías Imc Normal y Sobrepeso, de la variable *ObesidadOr*, pero sin que aparezca la categoría Obesidad, primero tendremos que seleccionar casos mediante un filtro para quedarnos con ambas categorías. Se puede hacer con el cuadro *Seleccionar casos* optando por Si se satisface la condición de que *ObesidadOr* no sea la categoría 2 (Obesidad): `ObesidadOr ~= 2`.

Videotutorial 5A2Seleccionar.avi
Se muestra como se hace una selección temporal de los casos fumadores y obesos con el cuadro *Seleccionar casos*, con la finalidad de realizar la estadística descriptiva y el diagrama de caja de la Glucosa de los fumadores y obesos. Después se deshace la selección para posteriormente hacer una selección permanente de los casos procedentes del hospital Ramón y Cajal (el código del Ramón y Cajal es 1) que guardaremos en el archivo de datos *FactoresRyC.sav*.

Ordenación y Segmentación de archivos

El cuadro *Ordenar casos* permite ordenar los casos del archivo de datos según los valores de la variable o variables especificadas. Simplemente se deben seleccionar la o las variables y especificar el criterio de ordenación: ascendente (de menor a mayor) o descendente (de mayor a menor). Cuando se especifican varias variables los casos se ordenan en primer lugar por los valores de la primera variable, después por los de la segunda y así sucesivamente, y cada una puede emplearse en orden ascendente o descendente.

El cuadro *Segmentar archivo* segmenta la matriz de datos en subgrupos de casos que serán analizados separadamente cuando se realice cualquier procedimiento estadístico. La matriz de datos debe estar ordenada por la o las variables de segmentación elegidas, por lo que si no lo está, se debe de ordenar previamente. Si queremos conocer la distribución de frecuencias de la variable Sexo en cada uno de los cinco hospitales, primero el archivo debe estar ordenado por Hospital, después se segmenta el archivo por Hospital y finalmente se realiza la distribución de frecuencias de Sexo. Se obtienen los porcentajes de hombres y mujeres en cada una de los segmentos por lo que esté segmentado el archivo, cinco en este caso, uno por cada hospital.

Videotutorial 5A3Segmentar.avi
Se muestra como se ordena el archivo por las variables Sexo en orden descendente (de mayor a menor, primero 1=Varón y después 0=Mujer) y Edad en orden ascendente (de menor a mayor) con el cuadro *Ordenar casos*. También se muestra como se obtiene la distribución de frecuencias de Sexo en cada uno de los cinco hospitales mediante la segmentación del archivo con el cuadro *Segmentar archivo*. Finalmente se deshace la segmentación anterior.