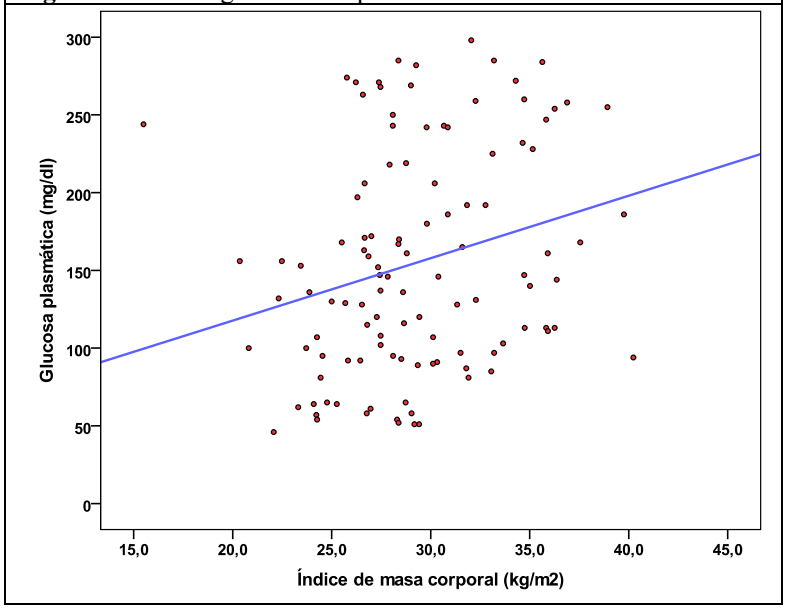
13A. RELACIÓN ENTRE DOS VARIABLES CUANTITATIVAS PRUEBAS ESTADÍSTICAS DE CONFORMIDAD

RELACIÓN ENTRE DOS VARIABLES CUANTITATIVAS

Las pruebas estadísticas para analizar la relación entre dos variables cuantitativas, *Glucosa* e *Imc* en este caso, son la *Correlación* y la *Regresión lineal simple*. Esta relación se explica y entiende mejor comenzando con la valoración de su representación gráfica conjunta mediante el *Diagrama de dispersión* o de puntos (*Scatterplot*).

- **Diagrama de dispersión**
  Se realiza con la variable respuesta en ordenadas y la variable exposición en abscisas (Figura 13A.1). La forma alargada y ascendente de la nube de puntos sugiere una asociación lineal positiva entre ambas variables. Al aumentar el índice de masa corporal, se obtienen valores mayores de glucosa. Se puede ajustar la *recta de regresión* por el método de mínimos cuadrados, obteniendo la recta que minimiza la suma de residuales al cuadrado. Examinando la desviación de la recta de regresión respecto a la horizontal se puede analizar la relación lineal entre dos variables cuantitativas. Una recta horizontal o vertical indicaría que las variables no están relacionadas linealmente, mientras que la recta inclinada indicaría asociación lineal.

Figura 13A.1. Diagrama de dispersión *Glucosa – Imc*.



- **Regresión Lineal**
  Es una prueba estadística de homogeneidad en la que las variables juegan un papel asimétrico. Existe una variable respuesta (*Glucosa* en este caso) y otra u otras (permite introducir varias) variables exposición (*Imc* en este caso) que pueden ser cuantitativas o binarias. Es un modelo más general y que permite más aplicaciones que la Correlación. La regresión lineal se expone en detalle en el curso "Análisis de Supervivencia y Regresión Lineal, Logística y Cox".

- **Correlación**
  Las variables juegan un papel simétrico. No existe una variable exposición o independiente y otra respuesta o dependiente. Se trata por tanto de una prueba estadística de independencia. Sólo permite estudiar la posible asociación lineal entre las variables. Los índices más empleados para describir conjuntamente y valorar el grado de asociación lineal entre dos variables cuantitativas son el *coeficiente de correlación de Pearson*, basado en valores, y el *coeficientes de correlaciona ordinal de Spearman*, basado en ordenaciones. La *covarianza* es el índice básico del que deriva el coeficiente de correlación. En la Tabla 13A.1 se muestran sus valores en este ejemplo.

Tabla 13A.1. Covarianza y correlación en la relación *Glucosa – Imc*.

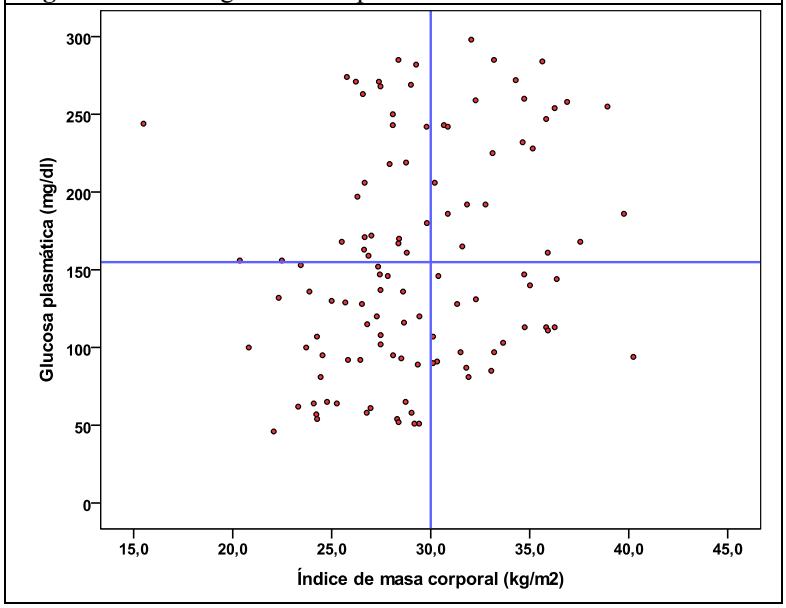
Prueba		**p**
Covarianza	79,3 kg/m^2*mg/dl	
Coeficiente de correlación de r Pearson	r = 0,250	0,008
Coeficiente de correlación de rho Spearman	ρ = 0,258	0,006

- **Covarianza** (s_{xy})

  Es la medida básica de variabilidad conjunta de dos variables. La covarianza entre dos variables cuantitativas X e Y (s_{xy}) es igual a la Suma de Productos cruzados entre ambas variables [$SP_{xy}=(x_i-M_x)*(y_i-M_y)$] dividida por los grados de libertad (n−1), es decir $s_{xy} = SP_{xy} / n-1$. La propiedad básica de la suma de productos cruzados es tomar valores positivos cuando la asociación lineal entre las dos variables es creciente y valores negativos cuando es decreciente. A partir de las líneas horizontal y vertical que pasen por la media de cada variable, el diagrama de puntos se ha dividido en cuatro cuadrantes que explican bien el signo del numerador de la covarianza y del coeficiente de correlación (Figura 13A.2). Los cuadrantes superior derecho (numerador de la covarianza es "positivo x positivo") e inferior izquierdo (idem "negativo x negativo") dan resultado positivo, mientras que los cuadrantes superior izquierdo (idem "negativo x positivo") e inferior derecho (idem "positivo x negativo") dan resultado negativo. Además el valor de cada producto será máximo en los puntos situados en la diagonal y mínimo en los puntos próximos a las medias.

  En el ejemplo, s_{xy} = 79,3 kg/m^2*mg/dl (Tabla 13A.1). Su valor depende de las unidades de medida lo que dificulta la valoración del grado de asociación lineal y la comparación de la asociación lineal entre dos parejas de variables.

Figura 13A.2. Diagrama de dispersión *Glucosa – Imc*.



- **Coeficiente de correlación r de Pearson**

  Es la covarianza estandarizada, desprovista de unidades de medida que se obtiene dividiendo la covarianza entre las desviaciones estándar de ambas variables ($r_{xy}=s_{xy}/s_xs_y=SP_{xy}/\sqrt{SS_xSS_y}$). Es un indicador del grado de asociación lineal entre las dos variables cuyo valor oscila entre −1 (asociación lineal negativa perfecta) y +1 (asociación lineal positiva perfecta) pasando por 0 (ausencia total de asociación lineal).

  En el ejemplo, r = 0,250 (p = 0,008) (Tabla 13A.1). Indica una aceptable asociación lineal directa: cuanto mayor es el *Imc*, mayor es la *Glucosa*. La nube de puntos del diagrama de dispersión es ascendente.

Para poder ser aplicado se requiere que la nube de puntos tenga una forma más o menos elíptica (en el ejemplo se puede asumir) y que las dos variables cuantitativas sigan leyes normales (Tabla 13A.2 y Figura 14A.4). En el ejemplo, las pruebas de normalidad de Kolmogorov-Smirnov con la corrección de Lilliefors (muestras grandes, > 30) son significativas, por mucho para *Glucosa* (p = 0,006) y por poco para *Imc* (p = 0,048).

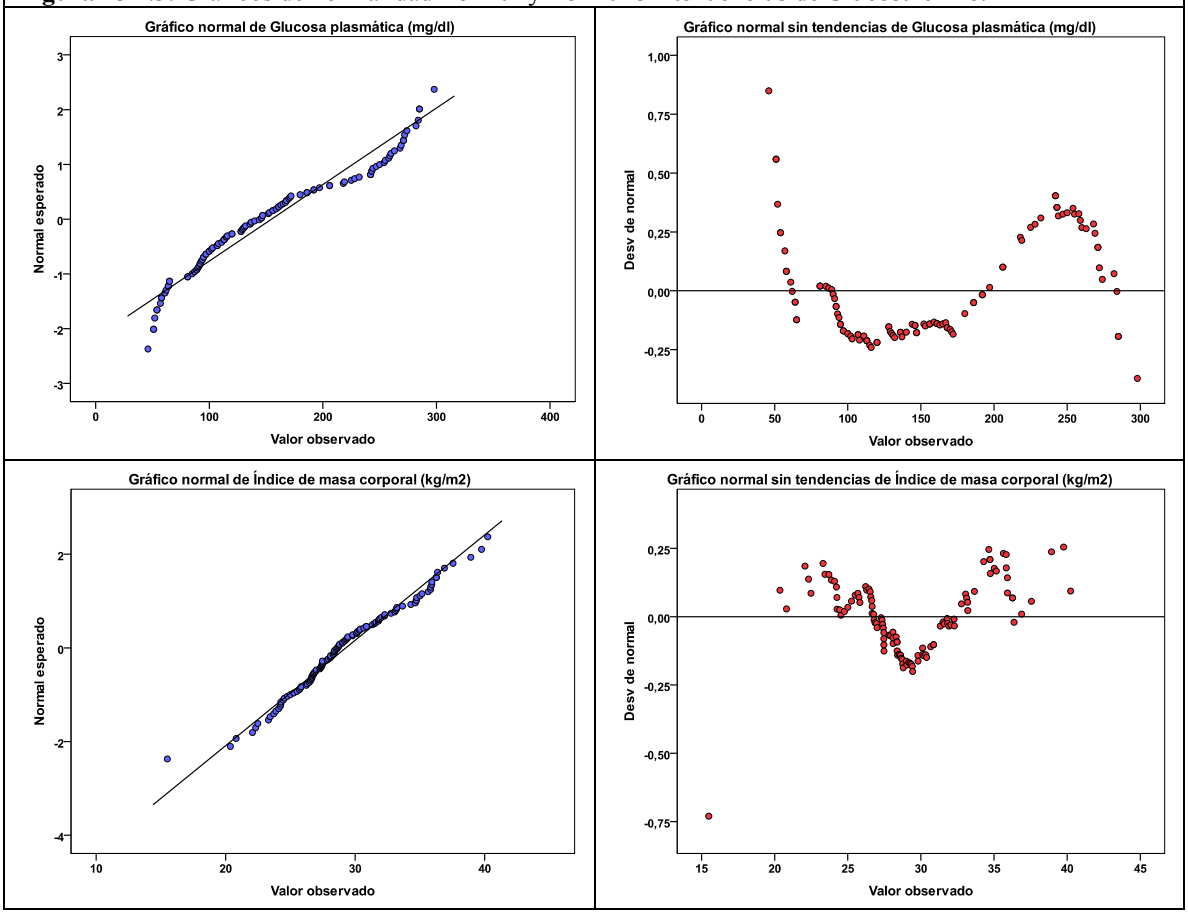
Tabla 13A.2. Pruebas de normalidad de *Glucosa* e *Imc*.

Prueba	Estadístico	gl	p
Kolmogorov-Smirnov con la corrección de Lilliefors de *Glucosa*	0,102	**112**	**0,006**
Shapiro-Wilks de *Glucosa*	0,938	112	<0,001
Kolmogorov-Smirnov con la corrección de Lilliefors de *Imc*	0,084	**112**	**0,048**
Shapiro-Wilks de *Imc*	0,984	112	0,206

El gráfico *Normal* de *Imc* se ajusta más a la diagonal que el de *Glucosa* (Figura 13A.3). El gráfico *Normal sin tendencias* de *Glucosa* tiene valores iniciales muy por encima de la horizontal y valores finales menos por encima e incluso por debajo de la horizontal, lo que indica más asimetría que el de *Imc* (Figura 13A.3). La

inspección de los gráficos de normalidad también corrobora que no se puede utilizar el r de Pearson por no cumplir el supuesto de normalidad.

Figura 13A.3. Gráficos de normalidad *Normal* y *Normal sin tendencias* de *Glucosa* e *Imc*.



- **Coeficiente de correlación ordinal Rho de Spearman (ρ)**
  Se utiliza en vez del r de Pearson cuando las variables cuantitativas no cumplen las condiciones de aplicación o para variables ordinales (Figura 13A.4). Al igual que otras pruebas no paramétricas se basa en ordenaciones y no se ve afectado por los valores anómalos. Es más robusto que el de Pearson y al no precisar condiciones de aplicación se puede emplear siempre. En el ejemplo ρ = 0,258 (p = 0,006) (Tabla 13A.1).

Figura 13A.4. Correlación entre dos variables cuantitativas: *Glucosa – imc*.

Correlación lineal (entre *Glucosa* e *Imc*)	
Normalidad (de *Glucosa* y de *Imc*) Pruebas: Kolmogorov-Smirnov y Shapiro-Wilks Gráficos: de Normalidad, de Caja, Histograma.	
Nube de puntos elíptica Diagrama de dispersión	
Sí	No
Coeficiente de Correlación r de Pearson	**Coeficiente de Correlación rho de Spearman**

Videotutorial 13A1Correlación1.avi
Con el archivo *Factores.sav* se muestra como se evalúa la asociación lineal entre las variables cuantitativas *Imc* y *Glucosa*. Primero se analiza la normalidad de *Imc* y *Glucosa* con los gráficos y pruebas de normalidad del cuadro *Explorar*. Después se realiza el gráfico de dispersión con el cuadro *Diagrama de dispersión simple* para ver la forma de la nube de puntos y obtener la recta de regresión. Finalmente se obtiene la covarianza y los coeficientes de correlación r de Pearson y rho de Spearman con el cuadro *Correlaciones bivariadas*.

Coeficiente de correlación

– El valor de la covarianza depende de las unidades de medida de ambas variables, por lo que no sirve para comparar el grado de asociación lineal de varias parejas de variables. El r de Pearson es la estandarización de la covarianza haciéndola independiente de las unidades de medida, por lo que permite comparar el grado de asociación lineal de varias parejas de variables. La covarianza entre *Imc* y *Glucosa* es 79,3 kg/m^2*mg/dl y entre *Peso* y *Talla* es 87,7 kg*cm (Tabla 13A.3). Estos valores indicarían un grado de asociación lineal similar, pero están influidos por las unidades de medida. El r de Pearson de *Imc* y *Glucosa* es 0,250 y el de *Peso* y *Talla* es 0,589 (Tabla 13A.3). Al no estar influido por las unidades de medida se pone de manifiesto que la asociación lineal es mayor del doble que la indicada por las covarianzas.

Tabla 13A.3. Covarianza y r de Pearson entre Glucosa – Imc y Peso – Talla.

Relación	Covarianza	r de Pearson
Glucosa - Imc	79,3 kg/m^2*mg/dl	0,250
Peso - Talla	87,7 kg*cm	0,589

– La magnitud aceptable de un coeficiente de correlación depende del tipo de estudio. En el ámbito de las matemáticas o la física, un coeficiente de correlación menor de 0,4 es malo, entre 0,4 y 0,7 es aceptable y mayor de 0,7 es bueno. Pero en investigación en poblaciones humanas y en biología o sociología, en general la variabilidad es mucho mayor por la mayor complejidad y un coeficiente de correlación de 0,4 es bastante aceptable.

– No detecta relaciones curvilíneas que son crecientes en un intervalo y decrecientes en otro, pero se puede emplear con variables ordinales y puede valorar una asociación entre las variables más general que la lineal.

Usos incorrectos del coeficiente de correlación

En diversas ocasiones el coeficiente de correlación no refleja la asociación entre las variables y su uso es incorrecto:

– **Correlación en la que está implicado el tiempo**. Hay variables que cambian a lo largo del tiempo y presentan una relación espuria. Ejemplo: relación entre precio de gasolina y tasa de divorcios en España en los últimos años.

– **Relación entre una parte y el todo**. Cambio relacionado con el valor inicial. Con el peso inicial P0 y al mes de tratamiento P1, siempre se encontrará relación entre P0 y la variación de peso P0−P1 ya en esta última interviene el peso inicial. Si P0 y P1 son independientes la correlación entre P0 y P0−P1 es de aproximadamente 0,7. Otro ejemplo: Correlación entre Peso e Imc.

– **Utilizarlo como medida de concordancia**. Si se mide una variable con dos procedimientos diferentes podemos encontrar un r muy alto pero la concordancia o acuerdo puede ser muy baja. La concordancia se mide con el coeficiente de correlación intraclase (CCI) para variables cuantitativas y con el índice kappa para categóricas como se verá en la unidad 15.

– **Utilizar muestras muy sesgadas**. El coeficiente de correlación se puede modificar mucho, e incluso invertir su sentido, eliminando de la muestra casos muy influyentes o con mucho peso e inversamente introduciendo casos diferentes a los de la muestra.

PRUEBAS ESTADÍSTICAS DE CONFORMIDAD

Las pruebas de conformidad verifican una hipótesis sobre un determinado valor de un parámetro de la población:

– ***Comparación de una proporción observada a una teórica***: ver si la muestra procede de una población con una determinada proporción de un carácter, por ejemplo igual proporción de hombres y mujeres (50%).

– ***Comparación de una media observada a una teórica***: comprobar si la muestra procede de una población con una media determinada, por ejemplo media de *Imc* de 25 kg/m^2.

El problema de estos contrastes es que son poco potentes: si no son significativos, no puede considerarse demostrado que los datos de la muestra se ajusten al valor teórico. Es un aspecto muy importante para no cometer este error en las conclusiones de un artículo. Sin embargo si son significativas, sí que puede considerarse que los datos de la muestra no se ajustan al valor teórico.

Comparación de una proporción observada a una teórica

En ocasiones es necesario comprobar la hipótesis sobre una determinada proporción sobre una característica de una población. Por ejemplo si se desea comprobar si la distribución del sexo de la muestra procede de una población con un 50% de hombres y un 50% de mujeres, se debe realizar una prueba estadística de conformidad para valorar la discrepancia entre las proporciones observadas en la muestra y las esperadas bajo el supuesto de equiprobabilidad (50% de hombres y mujeres).

La Tabla 13A.4 ofrece los valores observados y esperados de *Sexo* en cada categoría y la diferencia entre ellos. En la muestra hay 44 mujeres y 68 hombres. En una población con un 50 % de hombres y mujeres deberíamos esperar 56 hombres y 56 mujeres, es decir hay 12 hombres de más y 12 mujeres de menos. La cuestión es: ¿cual es la probabilidad de obtener por el error aleatorio de muestreo una muestra de 44 mujeres y 68 hombres de una población con igual número de hombres y mujeres?

Tabla 13A.4. Comparación de una proporción observada a una teórica del 50%.

	Mujer	**Varón**	**Prueba chi-cuadrado**
Observado	44	68	$\chi^2 = 5{,}14$
Esperado	56	56	gl = 1
Diferencia	−12	12	p =0,023

Esta probabilidad la ofrece la ley de χ^2 y también se muestra en la Tabla 13A.4. Aparece el valor del estadístico de contraste 5,14, los grados de libertad (gl = 1) y la significación estadística que en este caso es de 0,023, que al ser significativo (< 0,05) permite rechazar la hipótesis nula y concluir que la muestra no procede de una población con igual proporción de hombres y mujeres. Las condiciones de aplicación de la prueba χ^2 son que todas las proporciones esperadas sean mayores que 5 y en este caso se cumple (ambas son 56).

Comparación de una media observada a una teórica

En ocasiones es necesario verificar la hipótesis de que la muestra procede de una población con una determinada media en una variable cuantitativa. Por ejemplo si nuestra muestra, con una media de Imc de 29,3 kg/m^2, procede de una población con una media de *Imc* de 25 kg/m^2 (Tabla 13A.5).

Tabla 13A.5. Comparación de la media de *Imc* a la media teórica de 25 kg/m^2.

Media de Imc Observada en la muestra	29,3 kg/m^2
Media de Imc teórica	25,0 kg/m^2

El análisis de estas comparación se realiza con la ***prueba t de Student para una muestra*** o con la ***prueba no paramétrica T de Wilcoxon*** según se cumpla o no la normalidad, respectivamente, como se muestra en la Figura 13A.5.

Figura 13A.5. Comparación de una media observada a una teórica (Imc = 25 kg/m^2).

Comparación media observada a teórica
Imc: 29,3 – 25 kg/m^2

Normalidad
(de la variable *Imc*)
Pruebas: Kolmogorov-Smirnov y Shapiro-Wilks
Gráficos: de Normalidad, de Caja, Histograma.

Sí	No
Prueba t de Student-Fisher para una muestra	**Prueba no paramétrica T de Wilcoxon**

- **Pruebas y gráficos de normalidad**
  Se debe evaluar si la variable *Imc* cumple la normalidad mediante las pruebas y gráficos de normalidad. Para una muestra de 112 casos la prueba de Kolmogorov-Smirnov con la corrección de Lillefors es la adecuada y resulta significativa por poco (p = 0,048; Tabla 13A.3). Por los gráficos de normalidad se podría asumir la normalidad: el Normal se ajusta bien a la diagonal y el Normal sin tendencias sugiere simetría (Figura 13A.3). Hay argumentos para concluir en ambos sentidos: que se cumple o que no se cumple la normalidad.

- **Prueba t de Student-Fisher para una muestra**
  Se utiliza cuando se cumple el supuesto de normalidad. En este caso es significativa (p < 0,001; Tabla 13A.6). Además ofrece una estimación de la magnitud de la diferencia respecto a la media teórica: 4,3 (IC95% 3,5 – 5,1) kg/m^2 (Tabla 13A.6). Se concluye que nuestra muestra no procede de una población con una media de Imc de 25 kg/m^2.

Tabla 13A.6. Pruebas estadísticas en la comparación de la media de Imc a la media teórica 25 kg/m^2

Prueba	Estadístico	gl	p	Diferencia de Medias (IC95%)
t Student para una muestra	t = 10,193	111	<0,001	4,3 (3,5 – 5,1) kg/m^2
T de Wilcoxon	Z = –7,821		<0,001	

- **Prueba no paramétrica T de Wilcoxon**
  Es la prueba no paramétrica para comparar variables cuantitativas apareadas en un diseño de medidas repetidas, que puede emplearse para comparar una media observada a una teórica en el caso de no cumplirse la normalidad. El resultado es significativo (p < 0,001; Tabla 13A.6) por lo que no se puede asumir que la muestra proceda de una población con media de Imc de 25 kg/m^2.

Videotutorial 13A2Conformidad1.avi

Con el archivo *Factores.sav* se muestra como se comprueba si la muestra procede de una población con una proporción de Varón/Mujer del 50% utilizando ***la prueba de chi-cuadrado*** con el cuadro *Prueba de chi-cuadrado*. También se analiza si la muestra procede de una población con una media de Imc de 25 kg/m^2 utilizando la ***prueba t de Student para una muestra*** con el cuadro *Prueba T para una muestra* y la ***prueba no paramétrica T de Wilcoxon*** con el cuadro *Pruebas para dos muestras relacionales*. Para utilizar la prueba T de Wilcoxon para este fin, primero se tiene que crear una variable constante con el valor del Imc teórico, 25 en este caso, y que llamaremos *Imc25*.