Manejo de Datos en Biología Computacional. Herramientas de Estadística: Trabajo Final

Fernando Leonel Da Rosa Jurao

Índice

1. Introducción **3**
1.1. Conjunto de datos 3

2. Desarrollo del trabajo **4**
2.1. Representación gráfica del dataset 4
2.2. Medidas características de centralización y dispersión 4
2.3. Intervalo de confianza y tamaño muestral 5
2.4. Distribución de los datos 6
2.5. Test de hipótesis 7
2.6. Análisis de dependencia de variables categóricas. 7
2.7. Ajuste de regresión lineal 8

3. Conclusiones **9**

1. Introducción

Actualmente, estoy realizando un doctorado en el Instituto LEICI de la Facultad de Ingeniería de la UNLP, involucrado en un proyecto de desarrollo de un dispositivo médico llamado Páncreas Artificial (PA). El PA tiene como objetivo regular la glucemia en personas con Diabetes Mellitus Tipo 1 (DMT1) y consta de un monitor continuo de glucosa (CGM), una bomba de insulina y un algoritmo de control. Mi trabajo se centra, principalmente, en el desarrollo de estos algoritmos utilizando diversas estrategias de control automático. La validación de estas estrategias es una parte esencial en el desarrollo de controladores para PA. Gracias al trabajo y avances de distintos grupos de investigación, se dispone de simuladores que representan la dinámica de la glucemia en personas con DMT1, además algunos de ellos utilizan modelos aprobados por la FDA que pueden sustituir los ensayos preclínicos en animales. Esto permite validar y evaluar diferentes técnicas de control mediante simulaciones en computadora, conocido como validación ïn silico".

1.1. Conjunto de datos

En el grupo de trabajo se desarrolló un algoritmo de PA llamado Automatic Regulation of Glucose (ARG), el cual fue validado experimentalmente en los primeros ensayos clínicos de Latinoamérica. Este algoritmo utiliza un controlador llamado LQG (Linear Quadratic Gaussian), que a pesar de haber obtenido resultados satisfactorios en dichos ensayos, también resulta compleja su implementación y su funcionamiento no es simple de entender para el personal médico que participa en las pruebas. Por esta razón, se exploró la alternativa de utilizar un controlador más sencillo de implementar y comprender llamado PD (Proportional Derivative). Para contrastar el funcionamiento del nuevo controlador (PD) con el original (LQG), se utizó un simulador en donde se creó un escenario virtual. En dicho escenario los pacientes (10 adultos) se encuentran bajo tratamiento y consumen una comida de 70 gramos de carbohidratos (gCHO) a la misma hora. Los datos que se presentan en este trabajo son los resultados de la evolución temporal de glucemia de cada paciente obtenida en dicho escenario con cada estrategia. Las muestras se toman cada un minuto.

En la Figura 1, se muestra la distribución de los datos en el Dataset. Las primeras diez columnas corresponden a los pacientes bajo el tratamiento "LQG", mientras que las últimas diez columnas corresponden al tratamiento "PD".

	lqg_adulto_1	lqg_adulto_2	lqg_adulto_3	lqg_adulto_4	lqg_adulto_5	lqg_adulto_6	lqg_adulto_7	lqg_adulto_8	lqg_adulto_9	lqg_adulto_10	pd_adulto_1	pd_adulto_2	pd_adulto_3	pd_adulto_4	pd_adulto_5	pd_adulto_6	pd_adulto_7	pd_adulto_8
0	110.000000	110.000000	110.000000	110.000000	110.000000	110.000000	110.000000	110.000000	110.000000	110.000000	110.000000	110.000000	110.000000	110.000000	110.000000	110.000000	110.000000	110.000000
1	110.000000	110.000000	110.000000	110.000000	110.000000	110.000000	110.000000	110.000000	110.000000	110.000000	110.000000	110.000000	110.000000	110.000000	110.000000	110.000000	110.000000	110.000000
2	109.999188	109.998190	109.999775	109.999875	109.999841	109.999980	109.999848	109.997535	109.999986	109.999985	109.999188	109.998190	109.999775	109.999875	109.999841	109.999980	109.999848	109.997535
3	109.997730	109.994985	109.999361	109.999641	109.999556	109.999944	109.999590	109.993323	109.999960	109.999958	109.997730	109.994985	109.999361	109.999641	109.999556	109.999944	109.999590	109.993323
4	109.995762	109.990710	109.998788	109.999309	109.999168	109.999895	109.999258	109.987874	109.999926	109.999921	109.995762	109.990710	109.998788	109.999309	109.999168	109.999895	109.999258	109.987874

Figura 1: Formato del dataset

2. Desarrollo del trabajo

2.1. Representación gráfica del dataset

En la Figura 2 se muestra el gráfico del valor medio y la desviación estándar de la glucemia de los pacientes separados por estrategia. En este caso se seleccionan valores de interés que aporten al análisis comparativo. Por un lado se extrae el valor máximo de glucemia luego de la comida, y por el otro un valor en ayuno.

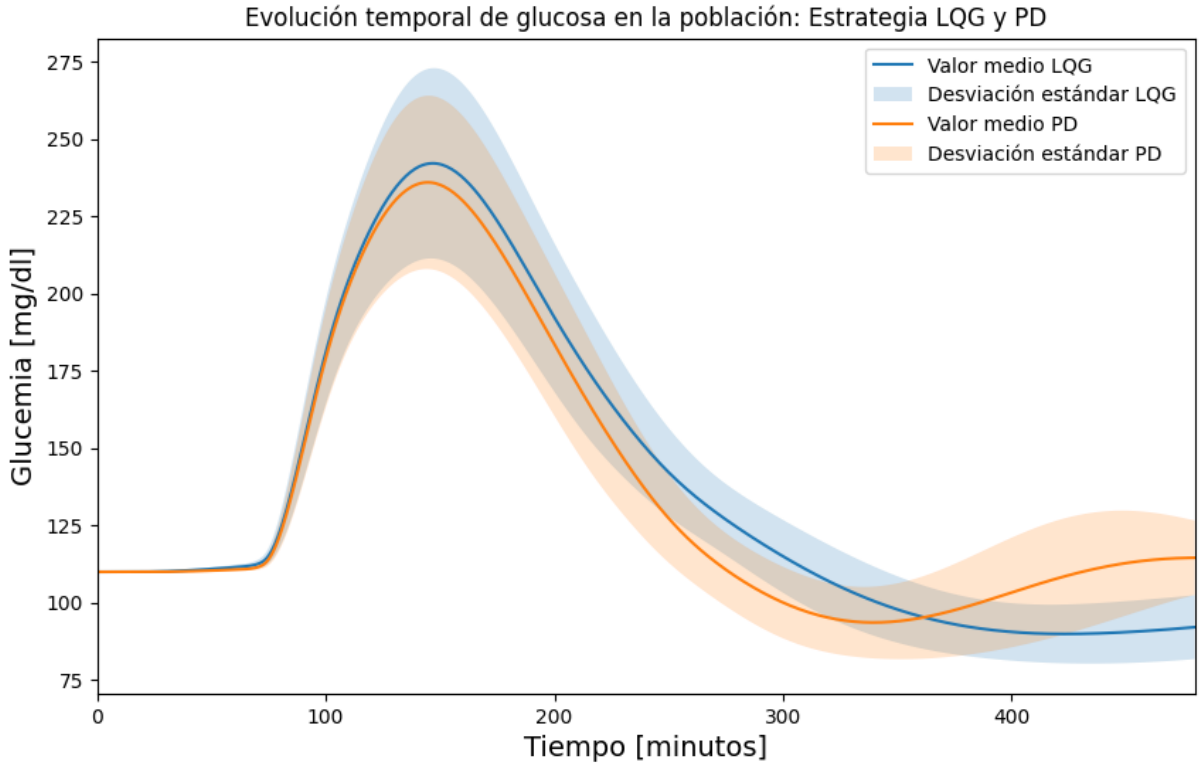


Figura 2: Evolución temporal de la glucemia. Se observa el valor medio y la desviación estándar de los 10 pacientes virtuales con cada estrategia (PD y LQG)

2.2. Medidas características de centralización y dispersión

Luego de extraer los valores de interés, se calcula el valor medio y la desviación estándar de ambos como medidas de tendencia central y dispersión de los datos. Los resultados se muestran en la siguiente Tabla.

	Valor medio [mg/dl]	Desviación estándar[mg/dl]
Glucemia máxima LQG	243.44	29.92
Glucemia en ayuno LQG	110.01	0.06
Glucemia máxima PD	237.45	27.15
Glucemia en ayuno PD	109.98	0.05

Tabla 1: Medidas de centralización y dispersión.

Según los datos de la Tabla 1, se puede observar que tanto la ubicación promedio de los valores como la variabilidad de los mismos son similares para ambos métodos. Sin

embargo, para poder obtener una conclusión lo mas acertada posible acerca de la similitud entre los métodos, se requiere realizar un estudio estadístico más adecuado.

2.3. Intervalo de confianza y tamaño muestral

Como primer paso, se estima promedio de cada valor para toda la población. Para ello se calcula el intervalo de confianza con un nivel de significancia $\alpha = 0{,}05$ (nivel de confianza $1 - \alpha = 0{,}95$). Al no conocer la varianza poblacional, se utiliza el test T de Student obteniendo lo siguiente:

- Valor máximo LQG: [175.75, 311.13] [mg/dl]
- Valor máximo PD: [176.04, 298.86] [mg/dl]
- Valor en ayuno LQG: [109.88, 110.14] [mg/dl]
- Valor en ayuno PD: [109.86, 110.08] [mg/dl]

Esto quiere decir que aproximadamente el 95 % de los intervalos obtenidos de forma repetida contendrán el valor promedio real de la población. Por lo que se puede decir que dicho valor se encuentra en el intervalo dado con un 95 % de confianza.

Si bien en este trabajo se dispone de un número limitado de muestras para el análisis, es de interés calcular el tamaño de la muestra necesario para obtener resultados estadísticamente significativos y confiables para el caso particular. Para esto se definen parámetros como el tamaño del efecto que se espera, la variabilidad del conjunto de datos, el nivel de confianza deseado.

Ya que el origen de los datos es una simulación, el calculo del tamaño del efecto no se basa únicamente en los propios datos, sino que se da un valor según lo que se espera observar. En primer lugar, en la Tabla 1 se observa que la desviación es similar entre pares de datos (ayuno y valor máximo) por lo que se utiliza la desviación total de los datos para calcular el tamaño del efecto. Además, al ser un análisis comparativo en donde se espera que el comportamiento sea similar entre métodos, se selecciona como valor esperable una diferencia de $1[mg/dl]$ en ayuno y $30[mg/dl]$ en el valor máximo. Esto último se debe a que el comportamiento frente a perturbaciones (en este caso la comida) de cada estrategia es distinto y por lo tanto se espera observar una diferencia mayor en el valor máximo que en el ayuno. En resumen, el tamaño del efecto queda determinado por la siguiente ecuación:

$$Effect\ Size\ = \frac{\Delta_{esperada}}{std_{conjunto}}$$

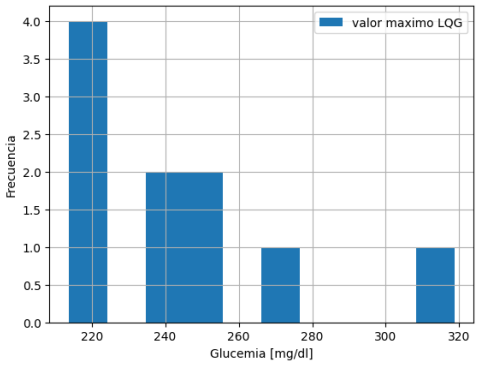
Seleccionando un nivel de significancia de $\alpha = 0{,}05$ y una potencia de 0.8, el tamaño muestral da:

- Para el valor máximo de glucemia: 15
- Para los valores en ayuno: 10

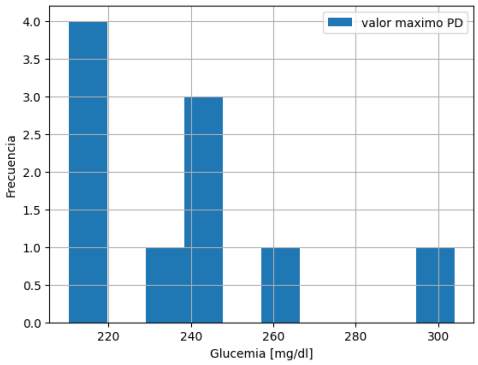
En el caso de los valores máximos, el tamaño muestral es superior al número de muestras del que se dispone (10 para cada método) por lo que en una situación ideal se tomarían más muestras. Esto quiere decir que las conclusiones que se obtengan de estos datos, no pueden ser generalizadas sin riesgo a cometer un error.

2.4. Distribución de los datos

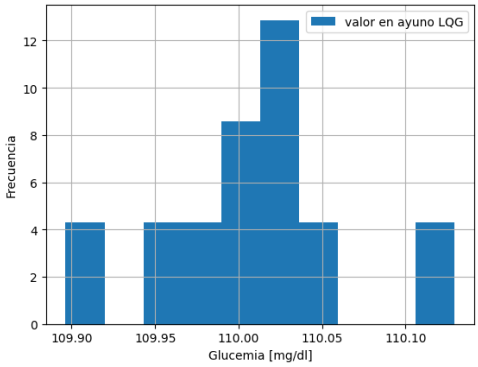
Antes de comenzar con el análisis comparativo se debe verificar si los datos siguen una distribución normal para luego decidir cual es el test que mejor se adapta en este caso. La Figura 3 muestra el histograma de cada parámetro. En principio, de lo observado en los gráficos, la conclusión es que los datos no siguen una distribución normal ya que se encuentran desplazados del centro. Aún así se realiza el test de normalidad para verificar. Debido a la poca cantidad de muestras se decide aplicar el test de Shapiro Wilk con la hipótesis nula H0 que dice que los datos siguen una distribución normal, y su alternativa H1 que dice que los datos no siguen una distribución normal. El resultado fue que tanto para el valor máximo de glucemia con el LQG como para el valor en ayuno con el PD se rechaza la hipótesis nula con una significancia menor a 0.05. Por lo que estos datos no se pueden considerar normales.



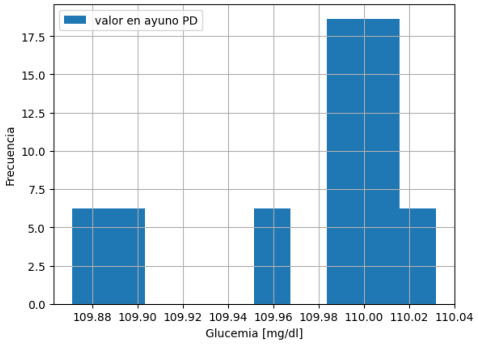
(a) Pico máximo LQG



(b) Pico máximo PD



(c) Ayuno LQG



(d) Ayuno PD

Figura 3: Distribución de los datos.

Otra medida de distribución de los datos, es la distribución de varianzas, para lo cual se utiliza el test de Levene con la hipótesis nula H0 de que los datos tienen la misma distribución de varianzas, y su alternativa H1 donde se dice que los datos tienen una distribución de varianzas distinta. En este caso no se rechaza la hipótesis nula, por lo que se acepta que las varianzas son similares.

2.5. Test de hipótesis

El objetivo con estos datos, es comparar si existe una diferencia entre utilizar un método u otro. Con el conocimiento que se tiene de los datos gracias a lo realizado en las secciones anteriores se puede seleccionar un test para la comparación. De los datos sabemos lo siguiente:

- Poseen varianzas semejantes entre grupos.
- Las muestras son pareadas ya que son tomadas de una simulación y por lo tanto los tiempos de observación son iguales para cada grupo.
- No siguen una distribución normal.
- Los mismos pertenecen a una variable continua.
- Las muestras están relacionadas ya que pertenecen a la misma población pero con distintos tratamientos.
- Las muestras son independientes.

Si bien no se cumplen los supuestos necesarios para aplicar un test paramétrico que asuma una distribución normal de los datos, en este caso si se cumplen para utilizar el test no paramétrico de Wilcoxon con las hipotesis:

- Hipótesis nula (H0): No hay diferencia significativa entre las medianas de los conjuntos.
- Hipótesis alternativa (H1): Existe una diferencia significativa entre las medianas de los conjuntos.

El resultado del test rechaza la hipótesis nula para ambos parámetros con un nivel de significancia menor a 0.05, lo que indica que hay evidencia estadística de que existe una diferencia significativa tanto en el pico máximo de glucemia como en el valor en ayuno entre la estrategia original y la nueva.

2.6. Análisis de dependencia de variables categóricas.

El dataset utilizado no contiene variables categóricas por lo que no es posible hacer el análisis con los datos de forma directa. Lo que se hace para realizar esta parte del trabajo es tomar muestras de cada paciente en un instante de tiempo (en este caso a las 5 horas) y luego agruparlas según la posición en la que se encuentren respecto al valor inicial de glucemia ($110[mg/dl]$) y al tratamiento al que correspondan. En la Figura 4 se muestran los valores tomados para el análisis.

Con las dos condiciones mencionadas se arma la tabla que se muestra a continuación.

	valor>110	valor<110
LQG	7	3
PD	3	7

Tabla 2: Tabla de contingencia

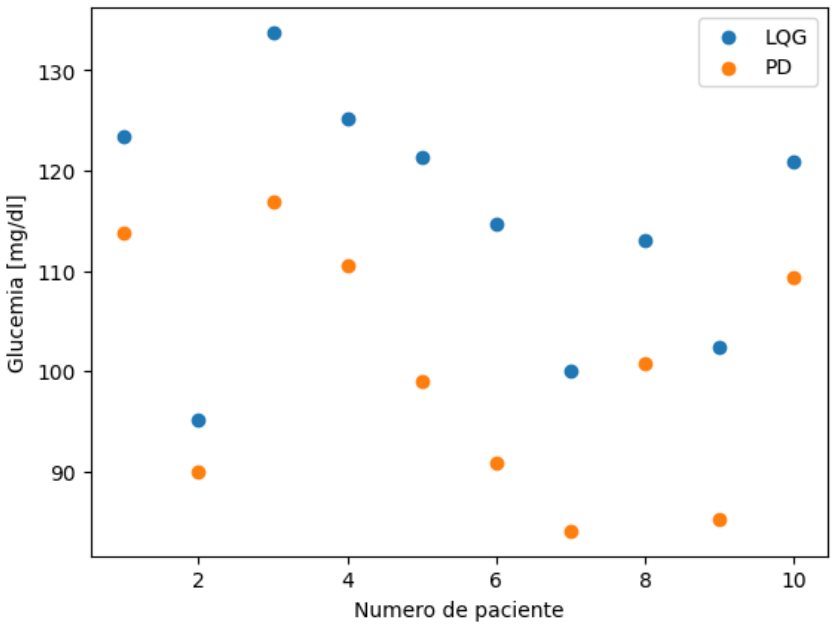


Figura 4: Datos para el análisis de dependencia de variables categóricas.

Donde se ve hay 7 casos en donde los que los valores superan los $110[mg/dl]$ y pertenecen al método LQG, 7 en donde están por debajo del umbral y pertecenen al PD, 3 casos del LQG que estan por debajo y 3 del PD que están por encima. Para determinar si hay una asociación significativa entre las dos variables, se utiliza el χ^2 (chi-cuadrado) test con las hipótesis:

- H0: no hay asociación significativa entre las dos variables categóricas en estudio.
- H1: existe una asociación entre las dos variables categóricas.

El resultado del test fue que no hay evidencia suficiente para rechazar la hipótesis nula (ρ *valor* $= 0{,}36$), por lo que se acepta dicha hipótesis y se dice que las variables categóricas no tienen relación entre sí.

2.7. Ajuste de regresión lineal

Para realizar esta parte del trabajo, fue necesario obtener datos de más de una simulación. Para cada una de ellas, se utilizó una comida con distinta cantidad de carbohidratos (hasta ahora se hizo el análisis con 70 gCHO). El objetivo es encontrar la relación entre la cantidad de carbohidratos de la comida y el pico promedio de glucemia máximo de la población, para ambos métodos.

Una vez realizadas todas las simulaciones en donde se usaron 50, 60, 70, 80 y 90 gCHO, se calcula el promedio del valor máximo en cada caso. Después, se utiliza el test de correlación de Pearson para ver si existe una relación lineal entre los valores máximos obtenidos y la cantidad de carbohidratos de la comida. En ambos casos se obtuvo un coeficiente de correlación de aproximadamente 1 con un nivel de significancia menor a 0.05 ($\rho = 0{,}0002$ para el PD y $\rho = 0{,}001$ para el LQG). Con estos resultados, se puede decir que la relacion entre la cantidad de carbohidratos y el pico máximo de glucemia es lineal. En la Figura 5 se muestran las correspondientes rectas de ajuste junto con los datos para cada método.

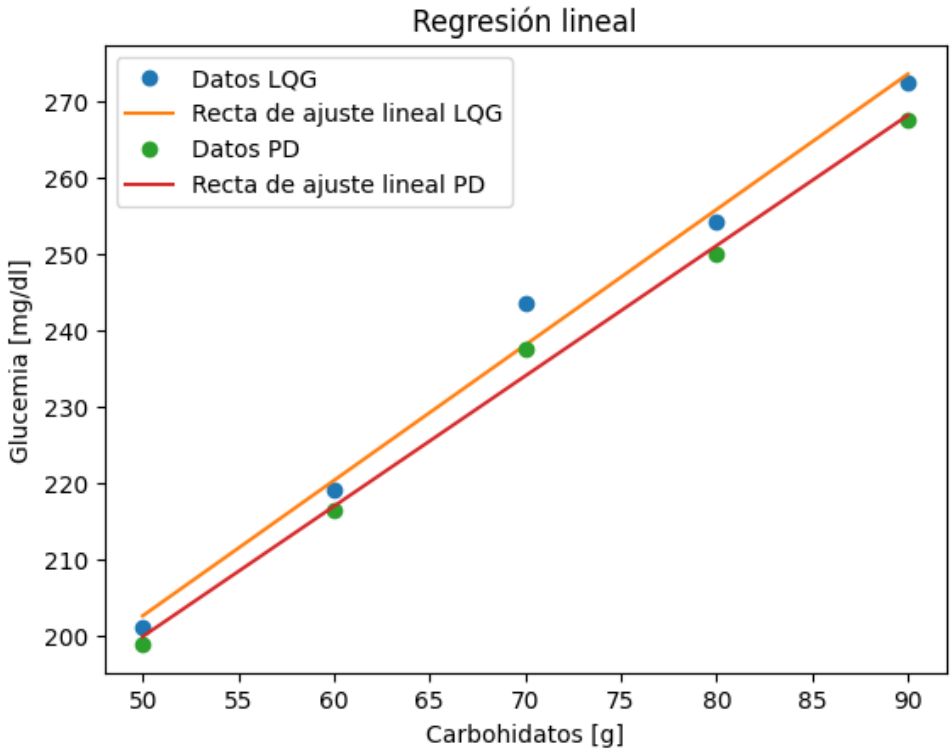


Figura 5: Recta de ajuste lineal para LQG y PD.

Aclaración: Es probable que la relación lineal que hay entre la cantidad de carbohidratos consumidos en la comida y el valor máximo de glucemia se deba a que la simulación realizada considera un escenario ideal y no incorpora algunas fuentes de incertidumbre que influyen en el resultado, generando un escenario más realista.

3. Conclusiones

En el trabajo se utilizó el lenguaje de programación "Python" para automatizar el procesamiento de datos y llevar a cabo el análisis estadístico. En particular, se realizó un análisis comparativo entre dos tratamientos para personas con Diabetes Tipo 1 y se llegó a la conclusión de que no son iguales en cuanto al efecto que tienen en la glucemia tanto en ayuno como al momento de compensar el efecto de la comida. Sin embargo, es importante destacar que esto no implica que un método sea mejor que otro, sino que para obtener una evaluación completa del funcionamiento de cada uno se deben seguir realizando pruebas.