# Prueba 1

**1.- Ordena de mayor a menor las series de precios por rentabilidad en el año 2003**

La rentabilidad de la serie dependerá del periodo de tiempo en que vayamos a invertir. Además imagino que la rentabilidad se calcula a toro pasado, es decir, sabiendo ya el precio final, lo que traducido a estadística es la probabilidad condicional sabiendo el resultado final por lo que se vuelve determinista.

Si solo nos centramos en inversión a un año. Entonces la rentabilidad de la serie de precios vendrá dada por:

$$R = \frac{Precio\ final\ -\ Precio\ Inicial}{Precio\ Inicial} * 100$$

Lo que viene siendo los "Returns" expresado en porcentaje.

Para este caso, el orden de rentabilidad de mayor a menor es: [1 5 3 4 2] con unas rentabilidades de [ 71.01654846 41.53057957 37.6912841 30.42217209 15.03428873] respectivamente.

**2.- Inventa una (o más) medida de 'riesgo' y ordena las cinco series con ella**

Probablemente que querréis una respuesta simple y corta pero como no tengo mucha experiencia en el campo ni entiendo muy bien lo que se me pide, establezco mejor primero lo que ahora mismo deduzco sobre el tema y luego ya respondo en concordancia.

El riesgo viene dado por la **incertidumbre** (distribución de probabilidad) que tenemos sobre el valor a estimar. Si el valor fuese predecible, no existiría riesgo. La incertidumbre viene dada por el "residual de nuestro modelo", es decir, la parte que no podemos predecir del mismo y que modelamos mediante una **variable aleatoria $\boldsymbol{e_r}$ .**

Nosotros intentaremos **construir una máquina** (modelo) que sea capaz de medir el precio en un instante $t$, $\boldsymbol{x_t}$, ha partir de toda la información disponible $\overline{\boldsymbol{X}}$ , como pueden ser su valor en instantes anteriores o el valor de otras señales con las que pueda tener relación, como pueden ser el historial de otros precios del mercado.

Al final tendremos un **vector de datos $\overline{\boldsymbol{X}}$** del cual suponemos que depende en cierta medida el precio a predecir $y = x_t$. Nuestro sistema intentará captar las relaciones deterministas entre $\overline{\boldsymbol{X}}$ e $y$, y si es posible (caso de estimador estadístico, o mejor aún, bayesiano) estimar la distribución de probabilidad que los relaciona. El residuo será el resultante error $\boldsymbol{e_r}$ a través del cual mediremos el riesgo.

$$x_t = f(\overline{\boldsymbol{X}}) + e_r$$

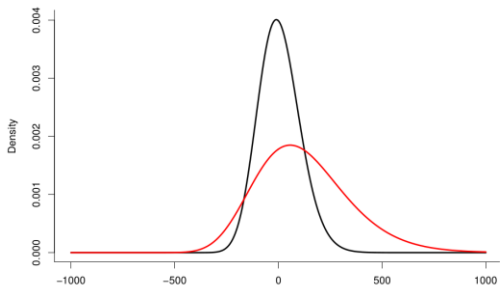
Este residual viene dado por varias componentes, tanto por información de la que no disponemos, como por ruido de muestreo, como por error en el estimador ya que nuestro $f(\overline{\boldsymbol{X}})$ no tiene porque perfectamente captar las relaciones latentes entre $\overline{\boldsymbol{X}}$ e $y$.

También tenemos que tener en cuenta si la función de distribución del residuo cambia con el tiempo o no, es decir, si el proceso que lo caracteriza es **estacionario en sentido amplio**, si bien, idealmente, el algoritmo de aprendizaje podría aprender en cierta medida esta dependencia con el tiempo e indicarnos para un momento dado los estadísticos del residuo en función de $t$.

Haciendo predicción de los returns y descomponiendo la señal (quitando por ejemplo componentes frecuenciales deterministas fuertes y trends polinómicos bajos) podemos en cierta medida hacer el residual más estacionario, sin embargo existirá probablemente cierta **heterocedasticidad** que también se podrá modelar con otros procesos como el GARCH (más que nada para modelarla y estimar la varianza) pero que no evitará que el riesgo dependa del instante de tiempo $t$.

Así pues, partiendo de esta base, que tampoco puede ser muy rigurosa porque no he trabajado mucho en finanzas ni en series temporales, no sé muy bien cómo caracterizar el riesgo de las 5 series de precios en función del estimador que se me ocurra a continuación. Y tampoco tengo mucho tiempo así que por lo menos en esta entrega no podré realizarlo. A lo mejor debería trabajar sobre los datos hasta quedarme con un residual decente y utilizar la distribución del residual como incertidumbre sobre la que calcular el riesgo, claro que esto conllevaría asumir que el residuo es estacionario, o bien si no lo es, estimar la incertidumbre en cada $x_t$ (modelo de proceso heterocedástico), estimar el riesgo sobre esta y después hacer sumatorio de los mismos. Sería interesante pero ahora mismo no tengo tiempo.

Vamos ahora con los **estimadores de riesgo**, dando por hecho que tenemos una función de distribución sobre la que calcularla y que está realizada sobre el **dominio de los Returns**. Supongamos también que la distribución es cóncava, es decir, que tiene un único máximo y que su media se encuentra cerca del punto de mayor probabilidad (cuasi simétrica). Dentro de esta asunción entran la mayor parte de distribuciones encontradas en la vida real, suma de gaussianas con la misma media, suma de laplacianas y otras distribuciones con colas más pesadas. La siguiente figura muestra 2 funciones de distribución que cumplen dichas propiedades.




A la hora de medir el riesgo también tendría que saber si **podemos tanto invertir como retirar acciones**. Es decir, si por ejemplo la distribución está muy desviada hacia las pérdidas, si podemos retirar acciones, es tan bueno como si lo hiciera hacia las ganancias,

en cuyo caso invertiríamos. De ser así entonces el riesgo se mediría aplicando el estimador al lado con más probabilidad, de no ser así, siempre se aplicaría a las pérdidas. A continuación aplicaremos los estimadores al segundo caso pero de poder des-invertir, el riesgo sería el mínimo entre los riesgos calculado para incremento y decremento de los precios.

Y por fin vamos con los estimadores:

1) **Varianza**. Dado que la varianza es un estadístico medidor de incertidumbre, es lógico utilizarlo para medir el riesgo.

2) **Varianza en el dominio para el cual Return < 0.** Podríamos afinar más y solo medir la varianza en el dominio de la distribución para la cual tendremos pérdidas. De esta manera, si la media de la pdf no está en 0, si no que está en las ganancias, midiendo la varianza de $-\infty\, a\, 0$, saldrá menor que la de la misma distribución y centrada en 0.

3) **Media en el dominio para el cual Return < 0 por la probabilidad de que Return < 0**. Establece una medida de las pérdidas esperadas por la probabilidad de que haya pérdidas.

**¿Opinas que el riesgo al escoger dos series depende exclusivamente del riesgo de cada uno de ellas por separado?**

Por supuesto que no, aún en el caso de que no estuvieran relacionadas (independientes), por el teorema central del límite, el resultado de la suma convergirá a una gaussiana con la media igual a la media de las distribuciones. Para el caso de 2 variables, la distribución resultante será la convolución de ambas. En el caso particular de que el estimador de riesgo sea lineal respecto a esta transformación entonces el riesgo sí que sería el mismo.

Si las variables están positivamente relacionada pues es probable que el riesgo aumente y si lo están negativamente (cuando sube el pan bajan los cubiertos) entonces el riesgo baja, aunque también lo hará la estimación del beneficio.

**4.- Inventa una (o más) maneras de hacer una predicción del precio de cada una de las series para el día 7-Jan-04.**

Básicamente a mi entender, aquí se están pidiendo técnicas a partir de las cuales modelas $f(\overline{\boldsymbol{X}})$ de tal manera que el residual resultante sea lo más determinista posible, es decir, estimaciones lo más certeras posibles de $x_t$.

$$x_t = f(\overline{\boldsymbol{X}}) + e_r$$

Podemos diferenciar en este caso 2 tipos de técnicas, las probabilísticas, que nos dan una función de distribución de $x_t$ y las que no. Siempre podremos utilizar de todas maneras

técnicas no probabilísticas para una primera estimación de $f(\bar{X})$ y después utilizar técnicas estadísticas sobre $y$' siendo

$$y' = \left(x_t - f(\bar{X})\right) + e_r$$

Donde $f(\bar{X})$ sea las relaciones que el primer sistema pudo captar. $\left(x_t - f(\bar{X})\right)$ por supuesto también tendrá una distribución estadística.

Podemos definir como $\bar{X}$ todos los instantes pasados de la señal hasta el momento $x_t$ a $x_0$. Y también los de otras señales de las que pueda depender, como son los precios de las otras señales hasta el momento. En vez de hacer la predicción utilizando todos los valores de instantes anteriores podemos definir una **ventana temporal de longitud $L$**, ya que eventos sucedidos hace mucho tiempo probablemente ya no condicionen el instante $t$, si ya sabemos los últimos $L$ instantes, lo que viene siendo propiedad markoviana en un sentido menos estricto. (Me imagino que no podrán utilizarse instantes futuros, pero de poderse, también los utilizaríamos ya que proporcionan información).

También, en vez de intentar predecir directamente $\bar{X}$, podemos predecir los **returns**, lo que elimina trends de tipo lineal y aplica un factor de escala, lo que facilita el aprendizaje. Sin embargo, también es posible que estemos perdiendo información al hacer esto. Por ejemplo, si el incremento de precio también depende de lo cerca que esté el precio de un cierto valor de saturación del mercado (por inventarme un contexto económico), el return no da información del precio actual. También podemos componer un vector formado por ambas informaciones (esto es algo exclusivo de las series temporales, normalmente no se puede hacer en machine learning).

Como es común en machine learning, habrá siempre un **conjunto de datos de training y otro de validación**, para comprobar que el modelo creado no hace overfitting, es decir, es muy preciso para los vectores de entrenamiento, pero muy malo para el resto. Normalmente es debido a que la complejidad del modelo es demasiado alta y este no generaliza, aprende las muestras exáctamente, aprendiendo el ruido que ello conlleva. Tan solo el tamaño de la ventana $L$ ya es un parámetro a validar ya que habrá valores que den mejor resultado que otros. Como medidas de performance se suele utilizar o bien el **MSE o el NLPD** (si el modelo es probabilístico).

**Modelos no estadísticos:**
Muchos de modelos aplicados en machine learning son no estadísticos y paramétricos, tampoco vamos a poner muchos, personalmente me he dedicado mucho más a la clasificación que a la regresión pero aquí van los que he utilizado:

- Regresión lineal.
- Redes Neuronales (y sus miles de variantes)
- Radial Basis Functions Networks.
- Group Method of Data Handling (y sus variantes GAME)

Por supuesto cualquiera de estos métodos podríamos utilizar en vez de la $\bar{X}$ inicial, versiones extendidas de la misma, mediante **expansiones** polinómicas, gaussianas, logarítmicas, exponenciales…

**Modelos estadísticos:**
Modelos que no solo ofrecen un valor puntual para estimar $y_t$ si no que también definen la función de distribución de la misma, tanto de f(X) como de e.

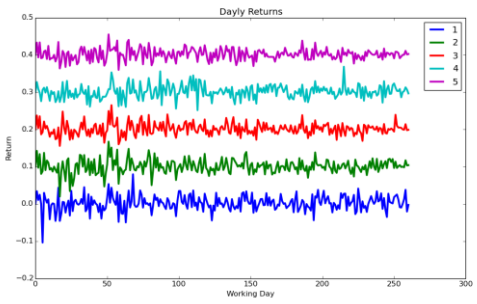
- Modelos **ARIMA** básicos: Que son simplemente establecer una relación lineal entre el precio actual y los anteriores (y también relación lineal exclusivamente con los ruidos anteriores). Filtro FIR ruidoso. El modelo también contempla hacerlo para varios niveles polinómicos del precio. Es decir, en vez de trabajar con el precio, trabajar con incrementos de precio (derivada) o con órdenes mayores.

- **Regresión lineal bayesiana**. Básicamente lo mismo que la regresión lineal pero aplicando una distribución de probabilidad a priori a los pesos w de la regresión y calcular su distribución a posteriori.

- **Proceso gaussiano.**
- Variaciones del proceso gaussiano para casos heterocedásticos.
- Mil variaciones del proceso gaussiano (e incluso laplaciano) que lo hacen el estado del arte hoy en día para regresión. Si no lo utilizan, lo recomiendo encarecidamente, aunque solo he trabajado con la versión básica.

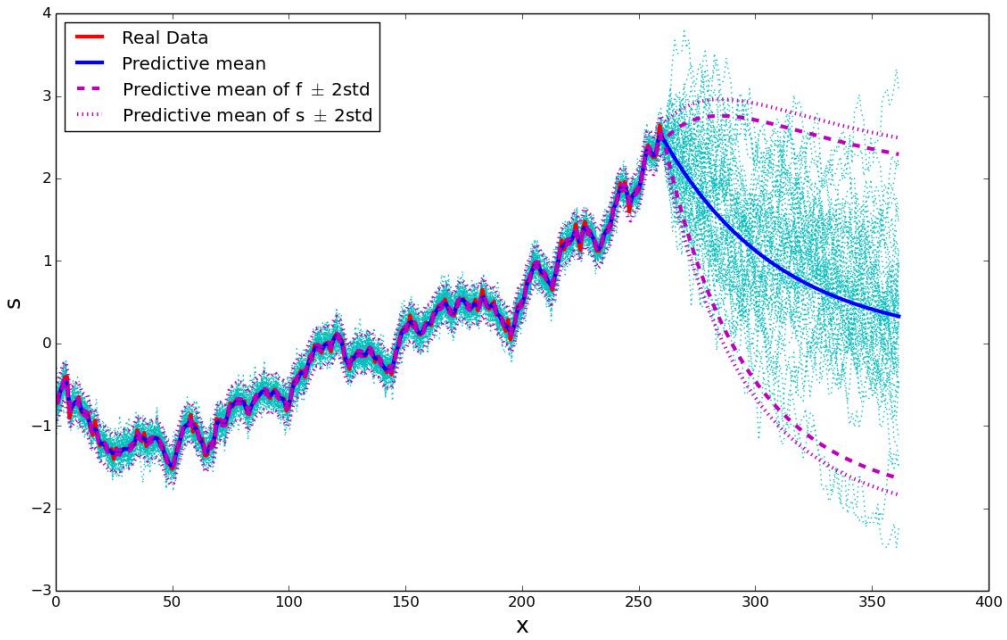
**5.- Por último, escoge en que dos series invertirías, y razona tu respuesta**

No he tenido tiempo a realizar estimaciones de las series de precios (aunque hubiera sido más asumir que los returns siguen un proceso estocástico estacionario y escoger las 2 series con más ganancia y menos incertidumbre que haber escrito todo lo que he hecho). Pero bueno, así a ojo (dado que los seres humanos somos las máquinas de análisis de datos más potentes del mundo, muchos sistemas del estado del arte funcionan con human rule-based methods) viendo las series de precios y los Returns:


Me quedo con las series 1 y 3. Claro que también pordría haber hecho estudios de dependencia entre las series, como la correlación, para decidir, por ejemplo las series 3 y 5 se ven bastante correladas.

Para demostrar que algo sí que he hecho aquí dejo una estimación bastante burda y a ojo de una de las series temporales utilizando la versión más básica del proceso gaussiano y calculando a ojo sus 3 parámetros $\sigma_0, \sigma_e \; y \; L$.




En rojo está la gráfica de los precios para una de las secuencias, habiendo normalizado los valores de los precios, en azul oscuro está la predicción del proceso gaussiano para el "medio día" de los mismos, es decir, predice para X = [0.5, 1.5, 2.5…]. Si bien estos instantes no existen, es solo una forma de ver las capacidades predictivas del algoritmo. Podemos ver que los días que cuando intentamos predecir el precio para días muy alejados de los datos, la varianza crece como es normal y el estimador de desploma a 0, que es prior del precio introducido. Cuando el punto a estimar está cerca de los datos, el precio depende de estos mediante correlación gaussiana de la nube de puntos a su alrededor (es bastante más complejo pero esa es la idea intuitiva). También están dibujadas varias realizaciones del proceso y los rangos de std tanto de $x_t$ como de la $f(\overline{\boldsymbol{X}})$ estimadas por el algoritmo.

# Prueba 2

**La terna 1487, 4817, 8147 es inusual en tres sentidos:**

**(i) cada uno de los tres elementos es primo**

**(ii) cada elemento es permutación de otro**

**(iii) es una progresión aritmética de constante 3330**

**No hay ternas análogas para primos de dos o tres dígitos, pero hay otra terna de cuatro dígitos. ¿Cuál es?**

Las ternas que mi programa ha encontrado son:

[0379, 3709, 7039]
[1487, 4817, 8147]
[2969, 6299, 9629]

Se podría argumentar que la primera también cumple las condiciones pedidas dado que siempre podemos expresar un número de menos dígitos, como uno de más dígitos, añadiendo ceros por la derecha.

El código básicamente primero calcula todos los números primos hasta el 9999 mediante un algoritmo de Sieves. Después, derivamos (con derivamos quiero decir derivé, pero siempre hay que escribir en plural o en impersonal) las posibles combinaciones de las ternas “abc” que son permutaciones y que se separan por “333”. En el código viene la derivación matemática ya que lo escribí allí en el momento. Los comentarios siempre los escribo en inglés por costumbre y buenos hábitos.

Tras el desarrollo, solo las inializaciones de “abc” que cumplen las propiedades:

$$a < 4, \quad b = a + 3, \quad c = a + 7$$
$$a < 4, \quad b = c + 3, \quad c = a + 4$$

pueden satisfacer las constraints. Lo que nos deja con solamente 8 posibles inicializaciones, que después buscamos entre los números primos precalculados y comprobamos que existan sus versiones incrementales de +3330 y +6660.

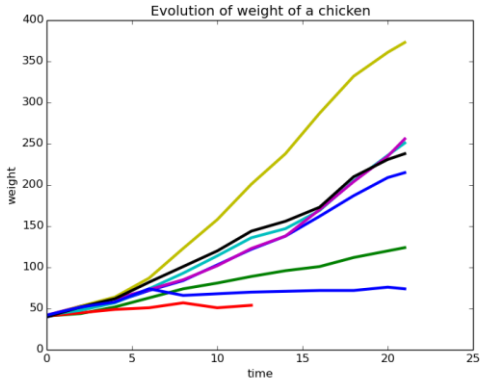
Como resultado tenemos las 3 ternas obtenidas.

# Prueba 3

**Responde de manera justificada a la pregunta: ¿Qué dieta hace que las gallinas ganen más peso?**

Para poder ordenar las dietas en función de la ganancia de peso de las mismas, primero necesitamos un estimador, es decir, un valor asociado a cada una de las dietas que caracterice la ganancia de peso. Tenemos la suerte de que todas las gallinas empiezan más o menos por el mismo peso lo que evitará sesgos en la estimación de la ganancia de peso debido a este factor.

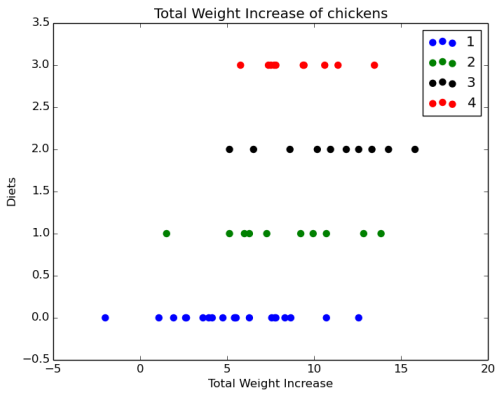
Conviene siempre antes de nada echar un vistazo a los datos, la siguiente gráfica muestra la evolución del peso de diferentes gallinas a lo largo del tiempo dado:




Se pueden ver gallinas que o bien no engordan o desaparecen antes de 21 días. Estos podrían ser outliers, pero como también podría darse que la dieta es demasiado agresiva o inexiste y mueran las gallinas no las podemos descartar sin un análisis más exhaustivo.

Se han tenido en cuenta 3 estimadores inventados para seleccionar la mejor dieta.
**1)** La ganancia de peso de cada gallina en función del número de días que sigue la dieta

$$E = E\left\{\frac{Ganancia\ de\ peso\ total}{Tiempo\ total\ de\ la\ dieta}\right\}$$



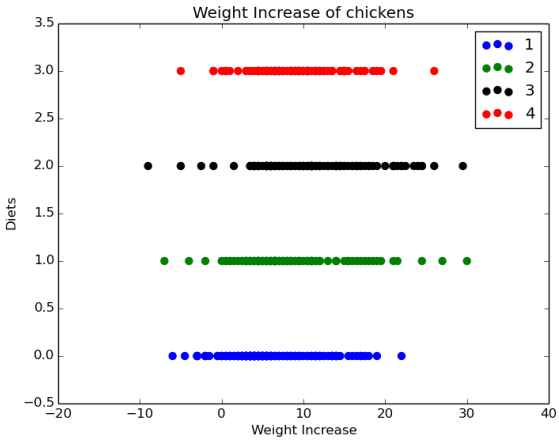

La gráfica de la derecha muestra el estimador para cada gallina de cada una de las 4 dietas. La media de la dieta 3 para este estimador es la mayor por lo que de acuerdo este estimador la dieta 3 sería la que hace que en media, las gallinas ganen el máximo peso de principio a fin de su dieta.

Dado que tenemos más gallinas en unas dietas que en otras, podríamos calcular también un estimador en base a intervalos de confianza, lo que favorecería a las dietas con más gallinas dado que ofrecen menos incertidumbre en la estimación. Sin embargo dada la clara superioridad de la dieta 3, el orden seguiría el mismo. El estimador utilizado no es

perfecto ni caracteriza totalmente la ganancia de peso ya que solo tiene en cuenta el principio y fin de la dieta.

**2)** Estimación de incrementos intervalo a intervalo de cada gallina.

$$E\ = E\left\{\frac{Ganancia\ de\ peso}{Tiempo}\right\}$$

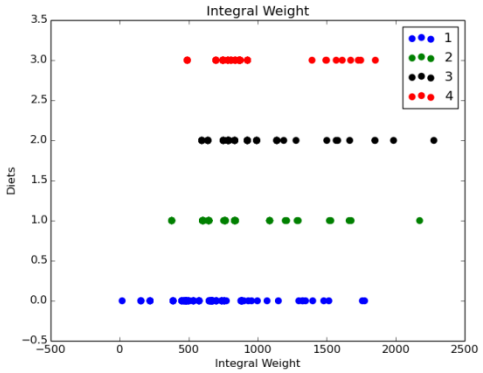



Como vemos, sale una versión más granulada del primer estimador. La nube de puntos obtiene cada vez más la forma de una gaussiana (Teorema Central del Limite). Nuevamente la dieta 3 es la que más engorda a las gallinas según este estimador. Sin embargo este estimador no valora positivamente la rapidez con la que engordan las gallinas, si se saturase el peso de una gallina en el día 10 por ejemplo, entonces el resto de incrementos serían 0, bajando la media.

**3)** Area bajo la curva del crecimiento de peso de la gallina / tiempo.

$$E\ = E\left\{\frac{1}{Tiempo\ total}\int_{Tiempo}(Peso - Peso\ inicial)\cdot dt\right\}$$

Si por ejemplo una dieta tarda tan solo 10 días en engordar al máximo la gallina y el resto su peso se satura (la gallina estructuralmente no puede engordar más), dicha dieta engorda más que otra que en los 21 días llegue al mismo resultado pero de forma lineal.




Vemos que la dieta 4 tiene 2 clusters por lo que la dieta hace que las gallinas ganen peso muy rapido al principio o al final, no tiene un incremento constante.

**Los 3 estimadores indican que la dieta número 3 es la que más hace engordar a las gallinas.**