Consejería de Agricultura y Pesca

Estudio del consumo y los precios al consumo de Frutas y Hortalizas

Aspectos Metodológicos

Marzo 2008

Versión 1

SECRETARÍA GENERAL DE AGRICULTURA, GANADERÍA Y DESARROLLO RURA

JUNTA DE ANDALUCIA

Estudio del consumo y los precios al consumo de Frutas y Hortalizas. Aspectos Metodológicos.

Índice de Contenidos

1. Introducción

A partir de la información de precios y consumo semanales proporcionados por TNS, se realiza un análisis con el objetivo de cuantificar las diferencias existentes entre los precios y/o consumos debidas a diferentes parámetros. Dado que la información ha sido extraída a partir de una muestra, las diferencias encontradas entre las medias de estos parámetros puede deberse al azar, por lo que se requiere un análisis estadístico de la información que permita determinar si las diferencias encontradas son significativas o si son debidas al propio proceso de toma de muestras.

En el presente documento se muestra la metodología llevada a cabo para estos análisis y que ha sido aplicada en la elaboración de estudios concretos para el tomate, el pimiento y el pepino.

2. Antecedentes

El análisis del consumo y de los precios al consumo que se expone a continuación se realiza a partir de la información proporcionada por TNS Worldpanel (empresa especializada estudios de mercado) y obtenida a partir del panel de consumo que realiza esta consultora. Este panel consta de 8.240 familias españolas, 1.494 de las cuales son andaluzas. A partir de dicho panel se obtiene semanalmente la cantidad consumida y el precio medio de un total de doce productos hortofrutícolas[1]. Esta información se encuentra desglosada según las siguientes variables:

- Localización: Andalucía y España (incluyendo Andalucía).
- Establecimiento: Descuento, Hipermercado, Supermercado y Tienda Tradicional.
- Presentación: Granel y Envasado.

De esta forma se obtiene, para cada producto y en cada semana, un total de 16 valores de precio y consumo semanal, correspondientes a todas las combinaciones posibles entre los valores de las variables enumerados anteriormente.

Además, se dispone de la información de precio y consumo total semanal por producto. Estos valores incluyen, además, otros establecimientos no contemplados anteriormente (cooperativas de consumidores, economatos, venta ambulante, etc.). Por lo tanto, los datos de consumo diferenciados a nivel de establecimiento no incluyen los consumos de este grupo de "otros establecimientos", ya que su importancia es reducida.

3. Metodología

Los valores medios resultan de gran interés para determinar diferencias entre los precios y los consumos debidas a diferentes factores. Sin embargo, la variabilidad propia de la información hace que existan diferencias entre estos valores, las cuales pueden ser reales o deberse tan

[1] Berenjena, Calabacín, Fresa, Judía Verde, Limón, Mandarina, Melón, Naranja, Pepino, Pimiento, Sandía y Tomate.

sólo al proceso intrínseco de la toma de muestras. Es por ello que se requiere realizar un análisis estadístico que permita establecer qué diferencias son significativas y cuales no lo son. A continuación se describen los procedimientos estadísticos que se abordarán en el estudio, así como el tratamiento efectuado a los datos.

El análisis estadístico de los datos se ha realizado mediante el programa SPSS 15.0 para Windows.

3.1. Análisis estadístico

3.1.1. Contraste de medias para muestras relacionadas

El contraste de medias para muestras relacionadas permite contrastar la hipótesis referida a la media entre dos muestras relacionadas.

Para este análisis, los valores se obtienen emparejados. Es decir, se toman n parejas de datos, que formarán las dos muestras. A partir de los valores obtenidos se calcula el estadístico T mediante la siguiente expresión:

$$T = \frac{\overline{X} - \overline{Y}}{\overline{S}_{\overline{X}-\overline{Y}}}$$

Donde:

$\overline{X}$: Media de la muestra X.

$\overline{Y}$: Media de la muestra Y.

$\overline{S}^2_{\overline{X}-\overline{Y}}$: Covarianza de las muestras X e Y.

De esta forma, este test equivale a realizar el test sobre la media de las diferencias, comparándolas con 0. En caso de que la media de las diferencias sea significativamente diferente de 0, las medias de las poblaciones X e Y serán significativamente diferentes, mientras que si se puede considerar que este valor es de 0, las medias de las poblaciones X e Y pueden considerarse iguales.

Para poblaciones normales, el estadístico T se distribuye según un modelo de probabilidad t de Student. En caso de que las poblaciones no sean normales, se puede asumir este mismo modelo de distribución si el tamaño muestral es suficientemente elevado (mayor a 20).

3.1.2. Tablas de contingencia. Comparación de proporciones

Cuando se trabaja con variables categóricas[2] (como es el caso que se está estudiando), los datos suelen presentarse en tablas de doble entrada en las que cada entrada representa una variable de clasificación. Como resultado de esta clasificación, las frecuencias (número de casos o porcentaje) aparecen organizadas en casillas que contienen información sobre la relación existente entre ambos criterios. A estas tablas de frecuencias se les llama tablas de contingencia.

[2] Las variables categóricas son aquellas cuyos valores son del tipo categórico, es decir, que indican categorías o son etiquetas alfanuméricas o "nombres".

En el caso que se está estudiando, en las tablas de contingencia que se construyen se sustituyen las frecuencias referidas al número de casos por la proporción de consumo, de forma que se muestra para cada combinación de las variables de clasificación empleadas, el porcentaje de consumo sobre el total del producto.

El grado de relación existente entre las dos variables no puede establecerse tan sólo observando las frecuencias de una tabla de contingencia. Para determinar si dos variables se encuentran relacionadas, se hace necesario utilizar alguna medida de asociación, acompañada de su correspondiente nivel de significación.

El estadístico que se emplea es el chi-cuadrado de Pearson, que permite contrastar la hipótesis de que los dos criterios de clasificación empleados (las dos variables categóricas seleccionadas) son independientes. Para ello, contrasta las frecuencias observadas con las esperadas (las que teóricamente se obtendrían si los dos criterios de clasificación son independientes). Cuando dos variables de clasificación son independientes, las frecuencias esperadas se obtienen mediante la siguiente expresión:

$$frecuencia\ esperada_{i,j} = \frac{total\ \ fila\ i \times total\ columna\ j}{n^{o}\ total\ casos}$$

Una vez obtenidas las frecuencias esperadas, se calcula el estadístico chi-cuadrado mediante la siguiente expresión:

$$X^2 = \sum_i \sum_j \frac{\left(n_{i,j} - m_{i,j}\right)^2}{m_{i,j}}$$

Donde $n_{i,j}$ se refiere a las frecuencias observadas y $m_{i,j}$ a las esperadas. De la expresión anterior se desprende que el estadístico chi-cuadrado valdrá cero cuando las variables sean completamente independientes (y por tanto las frecuencias observadas sean iguales a las esperadas), siendo mayor cuanto mayor sea la discrepancia entre las frecuencias observadas y esperadas. Este estadístico se distribuye según una distribución χ^2.

3.1.3. Análisis de la varianza. Comparaciones múltiples *post-hoc*

El análisis de la varianza de un factor sirve para comparar varios grupos en una variable cuantitativa. Se trata por tanto de una generalización de la prueba T para la comparación de medias.

La hipótesis que pone a prueba el análisis de la varianza es que las medias de las poblaciones estudiadas son iguales. Para poner a prueba esta hipótesis, se calcula el estadístico F, que refleja el grado de parecido entre las medias que se están comparando.

$$F = \frac{n\sigma^2_{\bar{Y}}}{\bar{S}^2_j}$$

Donde:

$n\sigma^2_{\bar{Y}}$: es un estimador de la varianza poblacional basada en la variabilidad existente entre las medias de cada grupo.

$\bar{S}_j^2$: es una estimación de la varianza poblacional basada en la variabilidad existente dentro de cada grupo.

Si las medias poblacionales son iguales, la estimación de la varianza poblacional basada en la variabilidad existente entre la medias reflejará el mismo grado de variación que la estimación basada en la variabilidad existente dentro de cada grupo, con lo que el valor de F será próximo a 1. Si las medias son diferentes, la primera estimación de la varianza reflejará un mayor grado de variación que la segunda, de forma que cuanto más diferentes sean las medias, mayor será el valor de F.

Si las poblaciones son normales y sus varianzas iguales, el estadístico F se distribuye según el modelo de probabilidad F de Fisher-Snedecor.

Por ello, previamente a este análisis se efectúa un contraste sobre igualdad de varianzas, empleando para ello el estadístico de Levene. En caso de que las varianzas sean diferentes, el estadístico F de Fisher-Snedecor se sustituye por el Brow-Forsythe, en el que el numerador es igual, sustituyendo el denominador por una estimación de la varianza de cada población por separado.

Estos estadísticos permiten tan sólo contrastar la hipótesis general de que los j promedios comparados son iguales. Al rechazar esta hipótesis se sabe que las medias poblacionales no son iguales, pero no permite conocer donde se encuentran estas diferencias, es decir, si son todas diferentes entre sí o si hay algunas que pueden considerarse iguales. Para determinar esto, se emplean una serie de contrastes denominados “comparaciones múltiples *post-hoc*”. Existen diferentes métodos que permiten realizar estas comparaciones. De todos los posibles, los utilizados mas frecuentemente son:

- Turkey, en caso de varianzas iguales.
- Games-Howel, en caso de varianzas diferentes.

3.1.4. Análisis de regresión lineal

Este tipo de análisis permite estudiar la existencia de relación lineal entre variables. Este análisis permite cuantificar el grado de relación existente entre dos variables, así como estimar una función matemática que describa la relación existente entre ambas, que en este caso tomará la forma de una recta:

$$y = a + b \times x$$

El objetivo de este análisis es doble:

- Determinar en qué medida la variable dependiente (x) puede estar explicada por la variable independiente (y).
- Obtener predicciones de la variable dependiente a partir de la independiente.

La forma de determinar los coeficientes a y b que conforman la recta de regresión es mediante el método de mínimos cuadrados. Es decir, los valores de a y b son aquellos que hacen que la suma de las distancias verticales entre cada punto y la recta (residuos) sea mínima. Estos valores permiten obtener las predicciones de la variable dependiente a partir de la independiente.

Para determinar en qué medida se relacionan estas variables se emplea el coeficiente de determinación (R^2), obtenido mediante la siguiente expresión:

$$R^2 = 1 - \frac{\text{Suma de los cuadrados de los residuos}}{\text{Suma de cuadrados total}}$$

Este coeficiente equivale al coeficiente de correlación de Pearson (r) elevado al cuadrado. El coeficiente de correlación de Pearson mide el grado de relación lineal existente entre dos variables. Toma valores entre –1 y 1. El valor 1 significa una relación lineal perfecta positiva, el valor –1 indica una relación lineal perfecta negativa y el valor 0 una relación lineal nula. Este coeficiente se calcula mediante la siguiente expresión:

$$r = \frac{\sum x_i y_i}{n S_x S_y}$$

Donde x_i e y_i representan cada una de las parejas de valores x e y, n es el número total de parejas de datos y S_x y S_y son las varianzas de las poblaciones x e y.

Resulta de interés contrastar si este coeficiente es significativamente diferente de 0, para lo cual se realiza un análisis de la varianza, en el que se compara la varianza debida al modelo y la debida a los residuos.

3.1.5. Nivel de significación

Según se ha visto anteriormente, los contrastes de hipótesis consisten en el cálculo de un estadístico, el cual se asume que tiene una distribución de probabilidad determinada. Posteriormente se calcula el nivel de significación asociado. Este parámetro representa la probabilidad de que las diferencias encontradas sean debidas al azar y se calcula a partir del valor del estadístico y la distribución probabilística del mismo (por ej., el nivel de significación del estadístico T se calcula a partir de la distribución t de Student, mientras que para el estadístico F se calcula a partir de la distribución F de Fisher-Snedecor).

Por lo tanto, cuanto menor sea el nivel de significación, es más probable que las diferencias no sean debidas al azar. Se establece como límite el 5% (0,05), de forma que si el nivel de significación supera este valor, se acepta la hipótesis de igualdad del parámetro estudiado (medias, varianzas o proporciones), y que las diferencias se deberán al azar, rechazándose esta hipótesis en caso de que el nivel de significación esté por debajo de este valor.

3.2. Tratamiento de los datos

Los datos de precios y consumo semanales se han dispuesto en una Base de Datos, lo que permite su agrupamiento a diferentes niveles.

En el caso del consumo, se obtiene el consumo del resto de España como la diferencia entre el consumo total de España y el de Andalucía. A la hora de comparar el consumo en ambos casos resulta necesario homogenizar los valores. Para ello, en lugar de comparar los valores de consumo totales se compararán los valores de consumo per cápita (kg/habitante), o bien la proporción del consumo según los diferentes valores que tome el factor estudiado. Con este objetivo, se le ha añadido a la BBDD la población total de Andalucía y del resto de España, obtenida del INE.

Por otro lado, se divide la campaña en cuatro periodos de 13 semanas cada uno. Esta división se realiza con el objetivo de poder estudiar de forma más sencilla la evolución de los precios y el consumo a lo largo de la campaña, comparando los valores medios de estas variables en cada periodo de tiempo. Los periodos considerados son los siguientes:

- **Periodo 1:** abarca desde la semana 36 a la 48 del año 2006. Se corresponde aproximadamente con el periodo otoñal.
- **Periodo 2:** abarca desde la semana 49 del año 2006 hasta la semana 9 de 2007. Se corresponde aproximadamente con el periodo invernal.
- **Periodo 3:** abarca desde la semana 10 a la 22 de 2007. Se corresponde aproximadamente con el periodo primaveral.
- **Periodo 4:** abarca desde la semana 23 a la 35 de 2007. Se corresponde aproximadamente con el periodo estival.