Atención oportuna a usuarios en salas de atención personalizada (estudio de caso)*

Revista Soluciones de Postgrado EIA, Número 2. p. 145-167 Medellín, junio de 2008

Diana Cecilia Uribe Cadavid**

* Artículo basado en el trabajo de grado exigido como requisito para obtener el título de Especialista en Gerencia de la Producción y el Servicio de la Escuela de Ingeniería de Antioquia. Director del proyecto: Juan Guillermo Villegas Ramírez

** Ingeniera de Producción, Especialista en Gerencia de la Producción y el Servicio, EIA, 2005. pfdianauribe@eia.edu.co

ATENCIÓN OPORTUNA A USUARIOS EN SALAS DE ATENCIÓN PERSONALIZADA
(Estudio de caso)

Diana Cecilia Uribe Cadavid

Resumen

Según la encuesta de satisfacción realizada en noviembre de 2004 por la empresa en estudio, se encontraron como factores de insatisfacción: tiempo de espera y atención del personal. Tiempo de espera fue la peor calificada. Se analizaron las causas de los tiempos prolongados y largas colas que estaban experimentando los usuarios. Se comprobó que la cantidad de clientes que llegaban al sistema de servicio no tenía ningún patrón de demanda que permitiera pronosticar épocas de mayor afluencia. Se analizó el sistema como un modelo de líneas de espera multicanal, una sola fase y se encontró que el número de clientes que llegaban al sistema en un período determinado se distribuía según Poisson, pero los tiempos de servicio no tenían el comportamiento exponencial requerido para solucionar matemáticamente el problema. Se procedió a simular el sistema con distribuciones para cada servidor, servicios de corta y larga duración, jornadas de descanso. Se evidenció que el sistema perdía capacidad durante la hora de almuerzo, aumentando la espera de los clientes. Se evaluaron soluciones para mejorar la capacidad.

Palabras clave: servicio, teoría de líneas de espera, simulación.

Abstract

According to the survey of satisfaction made by the company in study in November of 2004, they were like two dissatisfaction factors: time spent in queue and attention in charge of the personnel. Time spent in queue was the worst qualified. The causes of the long waiting times and large queues that the users were experiencing were analyzed. It was verified that the amount of customers who arrived at the system did not have any pattern of demand that allowed forecasting periods of greater affluence. The system was analyzed as a model of multi-channel lines with a single phase, where the number of customers arriving in a period has a Poisson distribution, but the service times did not have the exponential behavior required to solve mathematically the problem. The system was simulated using probabilistic distributions for each server and specifying the different services the customers needed. It was obvious that the system lost capacity during lunch time, increasing the time spent in queue of the customers. Several solutions were evaluated to improve the capacity.

Key words: service, queuing theory, simulation.

Atención oportuna a usuarios en salas de atención personalizada (estudio de caso)

Diana Cecilia Uribe Cadavid

Revista Soluciones de Postgrado EIA, Número 2. p. 145-167. Medellín, junio de 2008

Introducción

La investigación de operaciones se encarga de optimizar todas las operaciones que se llevan a cabo dentro de cualquier proceso productivo, permitiendo un mayor aprovechamiento de los recursos y, por lo tanto, una disminución en los costos.

El servicio al cliente se ha convertido en los últimos años en una ventaja competitiva más para las empresas, y la satisfacción de los usuarios, en un indicador de gestión obligatorio. El principio básico ahora es que "las empresas sobreviven gracias a sus clientes" y por esto ellos deben ser su objetivo principal.

La simulación es una herramienta que permite visualizar los resultados que tendrían las decisiones que parecen necesarias, con la ventaja de que ofrece no tener que hacer inversiones sin antes haber apreciado las consecuencias de cada una de las alternativas.

Las salas de atención personalizada tienen como misión prestar un servicio integral que permita a los usuarios satisfacer sus necesidades. Cuando de salud o de calidad de vida se trata, el tiempo se convierte en un elemento adicional del servicio y, por lo tanto debe controlarse para no generar insatisfacciones por largas esperas. Al estudiar las filas o líneas de espera de la empresa en cuestión, se pretende desarrollar una cultura de rápida atención y agilidad, que se convertirá en una ventaja competitiva de la organización frente a sus competidoras.

Para analizar entonces los tiempos de espera en cola de los usuarios de dichas salas, se deben encontrar el tipo de líneas de espera que tiene la empresa, las variables que las controlan y sus medidas de desempeño. Con esta información se pueden determinar las situaciones críticas y proponer opciones de solución para que sean evaluadas mediante simulación.

Metodología

Teoría de líneas de espera

Entender las de espera y cómo administrarlas es una de las áreas más importantes de la investigación de operaciones, ya que los modelos de colas o líneas de espera están en todas partes, todos los días. Todo hace cola: los aviones esperan para poder aterrizar o despegar, las personas esperan cuando entran a una estación de gasolina o llegan a una taquilla, los carros esperan para poder pagar el peaje...

Los tiempos de espera se han tornado cada vez más importantes debido al incremento en las exigencias de calidad en las operaciones de servicio, ya que los clientes relacionan la calidad del servicio con su velocidad. La mejor forma de disminuir los tiempos de espera es incrementando la capacidad de servicio, es decir, adicionando servidores al sistema, sin embargo, este aumento de la capacidad significa costos más altos de operación. Es aquí, entonces, donde la teoría de las líneas de espera cumple un papel fundamental, ya que su objetivo principal es encontrar el equilibrio entre el costo de mejorar el servicio (prestar un servicio más rápido) y el costo de hacer esperar a los clientes. La figura 1 muestra la relación de equilibrio esencial en condiciones de clientes (estables) típicas. Cuando la capacidad de servicio es mínima, el costo de la línea de espera es máximo, debido a los largos tiempos generados. A medida que aumenta la capacidad de servicio, se reduce el número de clientes en cola y, por tanto, los tiempos de espera, con lo cual disminuye el costo de la líneas de espera.

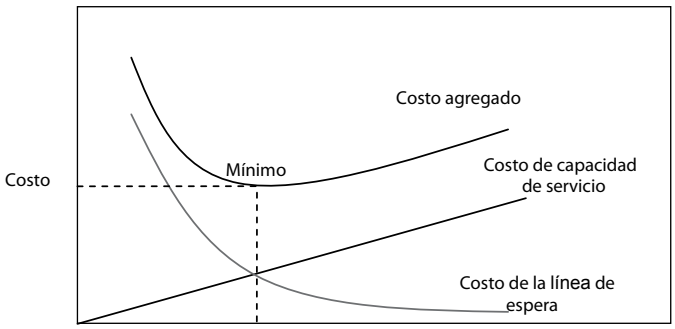

Figura 1. Costo contra capacidad de instalación de servicios

El origen de la teoría de línea de espera se basa en los problemas de congestión de las redes telefónicas y en el trabajo de A. K. Erlang (1878-1929), ingeniero y matemático danés, quien asesoró a la Compañía Telefónica de Copenhague (Russel y Taylor, 2000).

Una línea de espera se forma debido a que personas u objetos llegan al sistema de servicio más rápido de lo que pueden ser atendidos. Esto no indica necesariamente que el sistema de servicio carezca de suficiente capacidad para atender a sus clientes, sino que los clientes no llegan a un ritmo constante y los servidores no se demoran siempre lo mismo con todos los clientes. La teoría de líneas de espera, entonces, trabaja con las medias de las llegadas de los clientes y de los tiempos de servicio del sistema, ajustadas a ciertas distribuciones estadísticas (Russel *et al.*, 2000).

Figura 2. Elementos de las líneas de espera

- La población de clientes es la totalidad de clientes que pueden entrar a un sistema de servicio. Puede ser finita o infinita.
- La rata de llegada es la rata a la cual los clientes llegan al sistema de servicio, durante un período determinado. Las llegadas pueden ser descritas por varias distribuciones estadísticas, pero las soluciones analíticas para este tipo de problemas requieren que el tiempo entre llegadas se distribuya exponencialmente y que el número de llegadas por unidad de tiempo pueda definirse de acuerdo con una distribución de Poisson, con media λ (lambda) igual al número de llegadas en un tiempo determinado (Russel *et al.*, 2000).
- Tiempo de servicio es el tiempo requerido para atender a un cliente. También puede ajustarse a cualquier distribución estadística, pero la más usada en los modelos matemáticos es la distribución exponencial negativa. El servicio debe expresarse como la rata de clientes que son atendidos en un período μ.

Para que los modelos de líneas de espera tengan una solución promedio, la tasa a la cual los clientes son atendidos debe ser mayor que la tasa a la cual éstos van llegando ($\lambda < \mu$). De lo contrario, la cola crecería indefinidamente.

Tipos de líneas de espera

Los sistemas de líneas de espera generalmente se dividen en cuatro estructuras básicas de acuerdo con la naturaleza del servicio (figura 3) (Russel *et al*, 2000):

- Canal único, Fase única
- Canal único, Fases múltiples
- Múltiples canales, Fase única
- Múltiples canales, Fases múltiples

Los canales hacen referencia al número de servidores ubicados en paralelo para atender a los clientes que llegan al sistema y las fases se refieren al número de servidores secuenciales por los cuales los clientes deben pasar para completar el servicio.

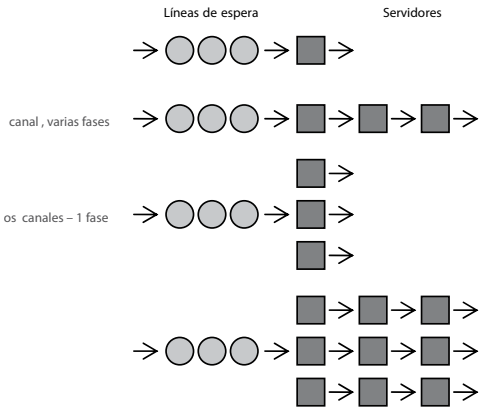

Figura 3. Estructuras básicas de las filas de espera.

Características de operación de las líneas de espera[1]

Las matemáticas usadas en la teoría de líneas de espera no proveen una solución óptima. Simplemente son características de operación del sistema en estudio que permiten evaluar su comportamiento y tomar decisiones. Se supone que estas características de operación serán constantes, una vez que el sistema haya alcanzado el estado estable. Las características de operación más usadas son (Chase, Jacobs y Aquilano, 2000):

L: número promedio de clientes en el sistema (en cola y atendidos)

L_q: número promedio de clientes en cola

W: tiempo promedio que los clientes esperan en el sistema (los en cola y los atendidos)

W_q: tiempo promedio de clientes que esperan en cola

P_0 probabilidad de que haya cero clientes en el sistema

P_n: probabilidad de que haya n clientes en el sistema

ρ: tasa de utilización, es decir, la proporción de tiempo que el sistema estuvo en uso

1 Para más información sobre líneas de espera ver Nico M. van Dijk, Why queuing never vanishes, European Journal of Operational Research, volume 99, issue 2, June 1997, páginas 463-476.

El modelo de líneas de espera que puede describir el comportamiento del sistema de servicio estudiado es el de múltiples canales, una sola fase, conocido también como (M/M/s). Las fórmulas utilizadas en este modelo para calcular las medidas de desempeño son:

$$P_0 = \left(\frac{1}{\left[\sum_{n=0}^{n=s-1} \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n \right] + \frac{1}{s!}\left(\frac{\lambda}{\mu}\right)^s \left(\frac{s\mu}{s\mu-\lambda}\right)} \right)$$

$$P_n = \begin{cases} \frac{1}{s!\,s^{n-s}}\left(\frac{\lambda}{\mu}\right)^n P_0, & para\ n > s \\ \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n P_0, & para\ n \le s \end{cases}$$

$$P_w = \frac{1}{s!}\left(\frac{\lambda}{\mu}\right)^s \frac{s\mu}{s\mu-\lambda} P_0 \qquad \rho = \frac{\lambda}{s\mu}$$

$$L = \frac{\lambda\mu \cdot (\lambda/\mu)^s}{(s-1)!(s\mu-\lambda)^2} P_0 + \frac{\lambda}{\mu} \qquad L_q = L - \frac{\lambda}{\mu}$$

$$W = \frac{L}{\lambda} \qquad W_q = W - \frac{1}{\mu}$$

Modelos de pronósticos de demanda

Un pronóstico es una predicción de lo que va a ocurrir en el futuro, por lo cual ayuda a las empresas a desarrollar más efectivamente sus operaciones y a tomar decisiones adecuadas frente a los problemas que se presentan en las empresas de servicios o manufactureras. Los directores de servicios están reconociendo la importancia de pronosticar en la mejora de la eficiencia y del servicio.

La selección de un sistema de pronósticos depende de tres factores (Chase *et al.*, 2000).

- Horizonte de planeación: los pronósticos pueden ser realizados a corto, mediano o largo plazo.
- Comportamiento de la demanda: dependiente o independiente.
- Patrones de demanda: demanda media, con tendencia, estacional, cíclica o aleatoria.

El patrón de demanda estacional se refiere a un movimiento oscilatorio de la demanda que ocurre periódicamente en el corto plazo y es repetitivo (Chase *et al.*, 2000). Por lo general, está relacionada con las estaciones, las horas pico, los períodos de vacaciones, las festividades especiales, etc. Los patrones estacionales pueden ocurrir mensual, semanal o diariamente. Para trabajar este tipo de demanda se utiliza un factor estacional, que es un valor numérico que multiplica el pronóstico normal de demanda para obtener un pronóstico

ajustado estacionalmente. El factor estacional es un valor entre cero y uno y es la porción del total de demanda asignada a cada estación. La mayoría de los servicios presentan patrón de demanda estacional, lo que lleva a una concentración de clientes en períodos específicos (Russel *et al.*, 2000).

Resultados de la encuesta de satisfacción

La encuesta se realizó a 2.172 personas, lo que corresponde al 0,61% de la población total (354.738 personas). Con esta muestra se logró un 95% de confianza y un 5% de error. Se evaluaron 8 salas de atención personalizada ubicadas en Medellín, Apartadó, Montería, Quibdó y Rionegro. Los resultados se encuentran en la tabla 1.

Tabla 1. Análisis de resultados encuesta de satisfacción al cliente

VARIABLE	INDICADOR (Medellín)	META	CUMPLIMIENTO
Horarios de atención	92,68%	90%	Cumple
Atención en la recepción	91,77%	90%	Cumple
Atención personal administrativo	82%	94%	No cumple
Tiempo de espera	67,16%	94%	No cumple
Comodidad en salas de espera	95,34%	94%	Cumple

La variable más crítica de las evaluadas es el tiempo de espera (67,16% de satisfacción), razón por la cual este trabajo se centró en el análisis de las causas de dicho problema.

Análisis de estacionalidad de la demanda

Se evaluó la existencia de diferentes patrones estacionales: mensual, semanal y diario, para encontrar si había algún mes, semana o día que tuviera mayor cantidad de clientes que estuvieran haciendo insuficiente la capacidad del sistema, generando largos tiempos de espera (figuras 4, 5 y 6).

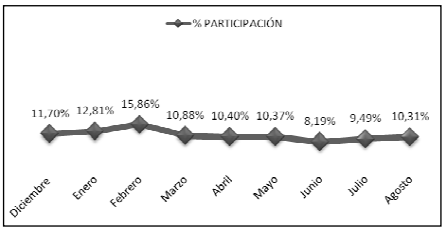

Figura 4. Turnos generados en la sala entre diciembre de 2004 y agosto de 2005

Los porcentajes de participación mensual muestran un comportamiento casi constante en la frecuencia de utilización del servicio en todos los meses, excepto febrero. El pico que se observa en ese mes se debe a dos factores identificados plenamente por la institución. Por lo tanto, se descarta la existencia de un patrón estacional mensual.

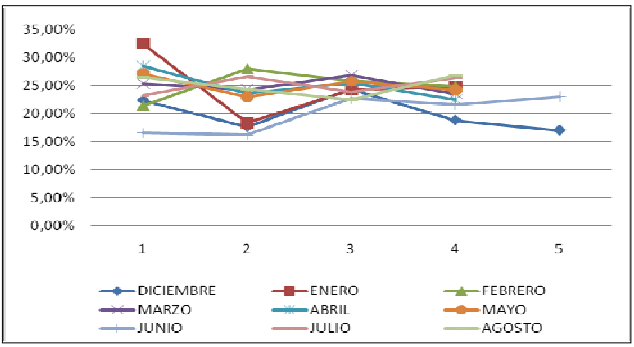

Figura 5. Turnos generados en la sala por semana en los meses de diciembre de 2004 a agosto de 2005

Los cálculos realizados de diciembre de 2004 a agosto de 2005 muestran que los porcentajes de participación de las diferentes semanas en el mes no tienen un comportamiento específico que pueda indicar la presencia de un patrón de demanda especial. Adicionalmente, no se observa una diferencia significativa entre los totales de cada semana que indique la existencia de un pico de demanda atribuible a alguna causa en particular. Por ello, se descarta la existencia de una estacionalidad semanal (anexo 1).

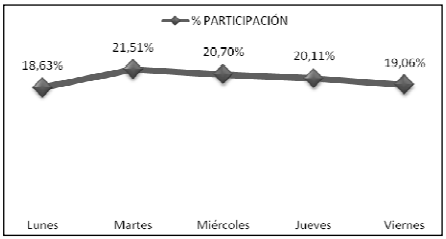

Figura 6. Turnos generados en la sala por día en los meses de diciembre de 2004 a agosto de 2005

Los porcentajes de participación obtenidos en el análisis de estacionalidad día a día tampoco muestran la existencia de picos de demanda. Por lo tanto, también se descarta la existencia de un patrón estacional diario.

La existencia de patrones de demanda estacional en las salas de atención personalizada obligaría a hacer ajustes transitorios en la capacidad de la sala. Al descartar la estacionalidad, sólo queda analizar la forma en la cual llegan los usuarios al sistema durante el día y así aplicar la teoría de líneas de espera para encontrar la razón por la cual los tiempos de espera son superiores a la meta de la empresa. Para aplicar correctamente la teoría de líneas de espera se debe analizar el tipo de sistema de servicio que se tiene, es decir, el número de canales y de fases del sistema.

Resultados aplicación del modelo de líneas de espera

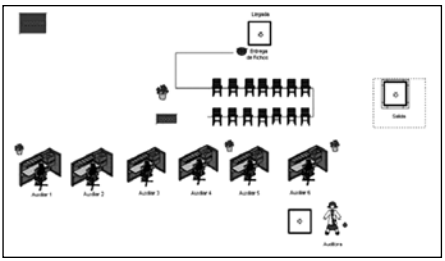

Figura 7. Sistema de líneas de espera actual simulado

El modelo de líneas de espera que es aplicable a este tipo de sistemas es el modelo (M/M/s). Las condiciones para aplicar el modelo son (Russel *et al.*, 2000):

- La rata de llegadas es Poisson
- El tiempo de servicio es exponencial
- La disciplina de la cola se rige por la regla FIFO
- La longitud de la cola es infinita y la población es infinita

Comportamiento de la rata de llegadas

Para hacer este análisis se tomaron todas las llegadas de los clientes al sistema durante dos días completos. Los tiempos entre llegadas se analizaron en el software Statgraphics, para confirmar si su comportamiento se regía por una distribución exponencial. Los resultados se muestran en la figura 8.

Figura 8. Media del tiempo entre llegadas

La prueba de ajuste realizada fue la chi-cuadrado. Los resultados para un 90% de confianza se pueden apreciar en la figura 9.

Goodness-of-Fit Tests for Col_1

Chi-Square Test

	Lower Limit	Upper Limit	Observed Frequency	Expected Frequency	Chi-Sq
at or below		0.313601	50	39.20	
	0.313601	0.664177	31	39.20	
	0.664177	1.06163	41	39.20	
	1.06163	1.52045	42	39.20	
	1.52045	2.06312	34	39.20	
	2.06312	2.7273	48	39.20	
	2.7273	3.58357	36	39.20	
	3.58357	4.79042	38	39.20	
	4.79042	6.85354	34	39.20	
above	6.85354		38	39.20	

Chi-Square = 8.66327 with 8 d.f. P-Value = 0.371494

Estimated Kolmogorov statistic DPLUS = 0.0383311
Estimated Kolmogorov statistic DMINUS = 0.0181415
Estimated overall statistic DN = 0.0383311
Approximate P-Value = 0.612178

EDF Statistic	Value	Modified Form	P-Value
Kolmogorov-Smirnov D	0.0383311	0.759603	>=0.10*
Anderson-Darling A^2	0.833979	0.835256	0.1882*

*Indicates that the P-Value has been compared to tables of critical values specially constructed for fitting the currently selected distribution. Other P-values are based on general tables and may be very conservative.

The StatAdvisor

This pane shows the results of tests run to determine whether Col_1 can be adequately modeled by an exponential distribution. The chi-square test divides the range of Col_1 into nonoverlapping intervals and compares the number of observations in each class to the number expected based on the fitted distribution. The Kolmogorov-Smirnov test computes the maximum distance between the cumulative distribution of Col_1 and the CDF of the fitted exponential distribution. In this case, the maximum distance is 0.0383311. The other EDF statistics compare the empirical distribution function to the fitted CDF in different ways.

Since the smallest P-value amongst the tests performed is greater than or equal to 0.10, we can not reject the idea that Col_1 comes from an exponential distribution with 90% or higher confidence.

Figura 9. Prueba de ajuste para el tiempo entre llegadas

Se confirma, entonces, que el tiempo entre llegadas para estos clientes sí tiene un comportamiento exponencial con media de 1,53 minutos entre llegadas.

Comportamiento del tiempo de servicio

Al analizar los datos de los tiempos de servicio de todos los servidores, se encuentra que el promedio es de 7,65 minutos por usuario. Al realizar la prueba de ajuste chi-cuadrado para encontrar la distribución a la cual se pueden ajustar los datos, se encuentra que la hipótesis de que los datos se distribuyen exponencialmente se puede rechazar con un 99% de confianza (figura 11). El ajuste a esta distribución no es posible, probablemente porque los tiempos de servicio de cada servidor siguen distribuciones de probabilidad diferentes. Esta hipótesis se validará más adelante.

```
Analysis Summary

Data variable: Col_1

372 values ranging from 1.0 to 35.0

Fitted exponential distribution:
     mean = 7.65323


The StatAdvisor
---------------
   This analysis shows the results of fitting an exponential
distribution to the data on Col_1.  The estimated parameters of the
fitted distribution are shown above.  You can test whether the
exponential distribution fits the data adequately by selecting
Goodness-of-Fit Tests from the list of Tabular Options.  You can also
assess visually how well the exponential distribution fits by
selecting Frequency Histogram from the list of Graphical Options.
Other options within the procedure allow you to compute and display
tail areas and critical values for the distribution.  To select a
different distribution, press the alternate mouse button and select
Analysis Options.
```

Figura 10. Media para el tiempo de servicio

```
Goodness-of-Fit Tests for Col_1

                              Chi-Square Test
-------------------------------------------------------------------------------
              Lower          Upper       Observed      Expected
              Limit          Limit      Frequency     Frequency    Chi-Square
-------------------------------------------------------------------------------
         at or below       0.806348             0         37.20         37.20
            0.806348        1.70777             5         37.20         27.87
             1.70777        2.72971            31         37.20          1.03
             2.72971        3.90946            45         37.20          1.64
             3.90946        5.30481            79         37.20         46.97
             5.30481        7.01258            66         37.20         22.30
             7.01258        9.21428            44         37.20          1.24
             9.21428        12.3174            45         37.20          1.64
             12.3174        17.6222            35         37.20          0.13
             17.6222           36.0            22         33.83          4.14
above           36.0                            0          3.37          3.37
-------------------------------------------------------------------------------
Chi-Square = 147.522 with 9 d.f.   P-Value = 0.0

Estimated Kolmogorov statistic DPLUS = 0.0655955
Estimated Kolmogorov statistic DMINUS = 0.227516
Estimated overall statistic DN = 0.227516
Approximate P-Value = 0.0

EDF Statistic          Value          Modified Form   P-Value
-------------------------------------------------------------------------------
Kolmogorov-Smirnov D   0.227516       4.4427          <0.01*
Anderson-Darling A^2   22.1991        22.235          2936.5756*
-------------------------------------------------------------------------------
*Indicates that the P-Value has been compared to tables of critical values
specially constructed for fitting the currently selected distribution.
Other P-values are based on general tables and may be very conservative.


The StatAdvisor
---------------
   This pane shows the results of tests run to determine whether Col_1
can be adequately modeled by an exponential distribution.  The
chi-square test divides the range of Col_1 into nonoverlapping
intervals and compares the number of observations in each class to the
number expected based on the fitted distribution.  The
Kolmogorov-Smirnov test computes the maximum distance between the
cumulative distribution of Col_1 and the CDF of the fitted exponential
distribution.  In this case, the maximum distance is 0.227516.  The
other EDF statistics compare the empirical distribution function to
the fitted CDF in different ways.
   Since the smallest P-value amongst the tests performed is less than
0.01, we can reject the idea that Col_1 comes from an exponential
distribution with 99% confidence.
```

Figura 11. Prueba de ajuste para el tiempo de servicio

Esta situación puede tener dos causas:

- Cada servidor atiende diez tipos de servicios diferentes. Cuatro de ellos requieren autorización del auditor de la sala. La empresa ha determinado unas duraciones ideales para cada tipo de servicio. Los promedios obtenidos por el sistema de información muestran que dichas metas son, en algunos casos, prácticamente inalcanzables. Por ejemplo, el servicio 3 actualmente tiene un tiempo promedio de atención de 9,50 minutos. La meta definida por la empresa para este servicio es de 8 minutos. La probabilidad

de alcanzar esta meta es realmente baja, ya que si el promedio de varios días es superior a la meta, significa que este servicio está teniendo muchos tiempos superiores y muy pocos inferiores a esa cifra.

Tabla 2. Tiempos promedio por servicio en la sala

SERVICIO	PROMEDIO DE SERVICIO
Servicio 1	8,00
Servicio 2	7,19
Servicio 3	9,50
Servicio 4	7,00
Servicio 5	5,31
Servicio 6	6,61
Servicio 7	9,65
Servicio 8	7,70

- La otra causa de la dispersión de los datos es que todos los servidores no trabajan con la misma eficiencia. La eficiencia depende en primer lugar de los métodos de trabajo que se empleen. En segundo lugar, la eficiencia es el resultado de la velocidad de los movimientos que efectúa el trabajador (Domínguez *et al.,* 1995). Para encontrar la distribución con la que se pueden modelar los tiempos de servicio de cada servidor, se realizaron pruebas de ajuste chi-cuadrado para los datos de cada uno de ellos (tabla 3).

Debido a estos dos problemas, el modelo de líneas de espera M/M/s no puede aplicarse, porque requiere que todos los tiempos de servicio se distribuyan exponencialmente con la misma media.

Simulación del sistema de servicio

El sistema de servicio se simuló utilizando el software de simulación discreta ProModel. Para simular más certeramente el sistema, se debe analizar de manera independiente cada uno de los servidores de la sala. El estudio se realizó con 5 servidores. La información registrada fue:

- La hora de generación del ficho
- La hora de atención en la taquilla
- La hora en la que el servidor tiene que desplazarse hasta donde está el auditor
- La hora de salida del cliente de la taquilla

El resumen de los resultados de los 5 servidores de acuerdo con la prueba de ajuste chi-cuadrado con un 90% de confianza se puede apreciar en la tabla 3.

Tabla 3. Resumen de los tiempos de servicio de cada servidor

Servidor	Cantidad de datos analizados	Promedio del tiempo de servicio	Desviación estándar del tiempo de servicio	Distribución a la cual se ajusta	Media	Desviación estándar	% de confianza
1	83	4,82	2,55	Lognormal	4,67	3,05	90%
2	44	13,61	11,93	Lognormal	13,89	14,22	90%
3	47	6,38	4,65	Lognormal	6,37	4,71	90%
4	91	5,42	3,57	Lognormal	5,4	3,6	90%
5	94	6,68	4,04	Lognormal	6,81	4,94	90%

Adicionalmente, se analizaron los tiempos de cada servicio para encontrar la distribución que mejor los representaba y se produjo la tabla 4.

Tabla 4. Distribución para servicios críticos

Servicio	Promedio del tiempo de servicio	Desviación estándar del tiempo de servicio	Distribución a la cual se ajusta	Media	Desviación estándar	% de confianza
2	6,07	5,51	Ninguna			99%
3	8,38	6,82	Ninguna			90%
4	7,0	3,8	Lognormal	7,0	3,8	90%
5	5,31	4,37	Lognormal	5,18	3,51	90%
6	6,61	3,77	Lognormal	6,61	3,77	90%
7	6,87	5,83	Ninguna			99%
8	7,7	8,19	Lognormal	7,9	8,41	90%
Mixtos	10,36	5,1	Lognormal	10,38	5,11	90%

Servicio	Mínimo	Máximo	Promedio	Moda
3	2	33	8,38	3
2	1	49	6,07	6
7	1	38	6,87	5

De todos los 359 clientes analizados, el 20% requirió la participación del auditor. El promedio del tiempo de atención del auditor es de 5 minutos con una desviación estándar de 1 minuto. Es decir, la variabilidad en la atención del auditor es alta, puesto que depende de la complejidad de la consulta.

Con esta información se pueden realizar dos simulaciones del sistema de servicio:

- Experimento 1. Según lo explicado y los resultados de la tabla 3, se puede confirmar que los servidores trabajan con diferentes eficiencias (cada uno tiene su propia

distribución de tiempos de servicio), sin importar el tipo de servicio que atienden.

• Experimento 2. Por otro lado, con los datos de la tabla 4 se puede suponer que el tipo de servicio es el que define el tiempo que tarda la atención y que nada tiene que ver con la persona que lo atiende.

Tabla 5. Parámetros de simulación experimento 1 y 2[2]

SIMULACIÓN	
Tiempo total de simulación	9 horas
Calentamiento	0 horas
Nivel de confianza	95 %
Error admisible	7%
Número de replicaciones	100

Los resultados para los experimentos 1 y 2 se aprecian en la tabla 6.

Tabla 6. Resultados simulación experimento 1

Parámetros		Experimento 1
		Media
Número de clientes promedio en el sistema	L	22,01
Tiempo promedio de los clientes que esperan en el sistema	W	31,86
Tasa de utilización 1	R1	87,33%
Tasa de utilización 2	R2	85,59%
Tasa de utilización 3	R3	84,09%
Tasa de utilización 4	R4	82,62%
Tasa de utilización 5	R5	80,11%
Tasa de utilización 6	R6	79,11%
Número promedio de clientes en cola	Lq	16,01
Tiempo promedio de los clientes que esperan en cola	Wq	23,50
Tiempo promedio de servicio	μ	8,15

2 El número de replicaciones requeridas se calcula con la fórmula $n=\frac{[z_{\alpha/2}\ S]^2}{e}$; donde z: área bajo la distribución normal que relaciona el nivel de confianza requerido, s: desviación estándar de los datos, e: error máximo permitido. Harrel, Ghosh y Bowden. Simulating using ProModel. 3 ed., p. 201.

Ambas simulaciones son concluyentes. El sistema de servicio presenta una cola (Lq) de más de 20 personas en promedio, alcanzando sus niveles máximos en los períodos de almuerzo de los servidores. Esta cola no logra ser disminuida durante el resto de la tarde, lo cual favorece la insatisfacción de todos los usuarios que lleguen después de las 12 m., puesto que los tiempos promedio de espera superan la meta de 15 minutos establecida por la empresa (Wq) (figura 12).

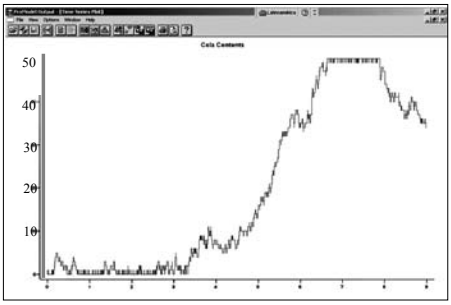

Figura 12. Comportamiento de la cola del sistema de servicio durante la jornada laboral

Se confirma, entonces, que tanto la variabilidad en la eficiencia de los servidores como la variedad en los servicios, afectan significativamente el desempeño del sistema de servicio.

Propuestas y resultados

Para disminuir los tiempos de espera de los usuarios de la sala y así alcanzar los indicadores de nivel de servicio, se pueden tomar varias alternativas.

- Propuesta 1. Cambiar la forma como salen los servidores a almorzar. Actualmente los servidores almuerzan en parejas. Desde las 12:30 p.m. comienzan a retirarse dos servidores del sistema por una hora. La figura 12 muestra que cuando comienza la hora de almuerzo, la cola empieza a crecer, y solamente se logra disminuir cuando la jornada de atención está terminando.

- Alternativa 1. Cambiar la forma de asignar el receso del almuerzo traslapando media hora la salida de cada servidor, es decir, cada media hora sale un solo servidor. De esta forma, el sistema está menos tiempo con menor capacidad. Esto puede afectar el clima organizacional, porque el primer servidor que almuerza y el último están más o menos media hora almorzando solos. Además, a todos los servidores se les trastorna un poco su horario de almuerzo.

- Alternativa 2. Mantener la asignación traslapada y disminuir el tiempo de almuerzo a 45 minutos. Con esto, se logra que la disminución de la capacidad de servicio del sistema sea más corta.

Los resultados de ambos experimentos se encuentran en la tabla 7.

Tabla 7. Resultados simulación propuesta 1, alternativas 1 y 2

		Actual	Alternativa 1	Alternativa 2
Número de replicaciones		100	200	200
Nivel de confianza		95%	95%	95%
Error admisible		7%	5%	5%
Parámetros		Media	Media	Media
Número de clientes promedio en el sistema	L	22,01	21,05	18,41
Tiempo promedio de los clientes que esperan en el sistema	W	31,86	30,05	25,51
Tasa de utilización 1	R1	87,33%	87,16%	89,95%
Tasa de utilización 2	R2	85,59%	85,32%	87,99%
Tasa de utilización 3	R3	84,09%	84,03%	86,64%
Tasa de utilización 4	R4	82,62%	82,09%	84,70%
Tasa de utilización 5	R5	80,11%	79,49%	82,10%
Tasa de utilización 6	R6	79,11%	78,43%	81,26%
Número promedio de clientes en cola	Lq	16,01	15,07	12,27
Tiempo promedio de los clientes que esperan en cola	Wq	23,50	21,91	17,38
Tiempo promedio de servicio	μ	8,15	8,05	8,04

En la tabla 7 se observa que los tiempos de espera en cola (Wq) y el número de clientes en cola (Lq) disminuyen cuando se traslapan y se reducen los tiempos de almuerzo. El efecto obtenido cuando sólo se traslapan los horarios de almuerzo no es estadísticamente significativo (anexo 2), mientras que el obtenido con la disminución del receso sí representa una mejora importante en el logro de la meta de los tiempos de espera en cola, ya que sus medias son estadísticamente diferentes. Sin embargo, para poder disminuir la hora de almuerzo debe hacerse una sensibilización con el personal para que entiendan dicha condición, y se les puede compensar con un descanso de 15 minutos en la tarde, que también sea traslapado, para garantizar que el sistema de servicio no se descompense.

- Propuesta 2. Adicionar un servidor. Se analizaron dos alternativas dentro de esta propuesta: añadir un servidor durante las 9 horas del turno y añadir el servidor solamente durante el rato normal de almuerzo de los servidores, es decir, 3 horas. Los resultados obtenidos para ambas opciones se muestran en la tabla 8.

Tabla 8. Resultados simulación propuesta 2, alternativas 1 y 2

		Actual	Alternativa 1	Alternativa 2
Número de replicaciones		100	500	500
Nivel de confianza		95%	95%	95%
Error admisible		7%	5%	5%
Parámetros		Media	Media	Media
Número de clientes promedio en el sistema	L	22,01	13,63	16,04
Tiempo promedio de los clientes que esperan en el sistema	W	31,86	18,53	21,94
Tasa de utilización 1	R1	87,33%	85,81%	86,92%
Tasa de utilización 2	R2	85,59%	82,62%	84,87%
Tasa de utilización 3	R3	84,09%	80,33%	83,31%
Tasa de utilización 4	R4	82,62%	77,15%	81,24%
Tasa de utilización 5	R5	80,11%	73,08%	78,37%
Tasa de utilización 6	R6	79,11%	71,59%	77,29%
Tasa de utilización 7	R7	N/A	67,75%	32,61%
Número promedio de clientes en cola	Lq	16,01	7,22	9,77
Tiempo promedio de los clientes que esperan en cola	Wq	23,50	10,02	13,63
Tiempo promedio de servicio	μ	8,15	8,19	8,09

En la tabla 8 se observa que los tiempos de espera en cola (Wq) y la cantidad de clientes en cola (Lq) disminuyen significativamente con ambas alternativas, pero la utilización de este séptimo servidor sería muy baja si trabajara todo el turno y, además, disminuiría la utilización (R) de todos los servidores. Por tanto, la propuesta de tener un servidor que apenas trabaje tres horas para compensar la pérdida de capacidad durante la hora de almuerzo es suficiente para lograr los tiempos promedio de espera de 15 minutos (anexo 2).

- Propuesta 3. Muchos de los tiempos que superan la meta establecida por la empresa requieren la intervención del auditor. Esto implica que el auxiliar de la sala deba desplazarse hasta donde el auditor y esperar a que él analice la situación, le dé la autorización para imprimir la orden, el auxiliar retorne a su puesto de trabajo, imprima la orden y regrese hasta donde el auditor para que le ponga su firma y sello. Este trámite no sólo es ineficiente por la cantidad de recorridos que implica, sino que también es muy demorado, porque el usuario está todo el tiempo parado en la taquilla esperando obtener una respuesta. Por lo demás, después de la espera, la respuesta puede ser negativa. Se proponen entonces cuatro soluciones:

- Opción 1. Que todos los servidores sean auditores. Esto permitiría al

mismo "auxiliar-auditor" autorizar todo tipo de servicios y dar las explicaciones pertinentes sin hacer esperar al usuario. Esta opción, aunque disminuiría los tiempos de espera, aumentaría enormemente el gasto administrativo de la sala por la diferencia salarial entre un auxiliar y un auditor.

- Opción 2. Desarrollar un sistema de comunicación interna que permita al auxiliar hacerle al auditor la consulta sin tener que moverse de su puesto de trabajo. El auditor ingresa a la historia del usuario y decide si autoriza o no el servicio. Si la respuesta es positiva, el software permite liberar la orden para que sea impresa por el auxiliar con la firma escaneada del auditor. Esta opción requiere un pequeño ajuste en el software de la empresa y también se debe desarrollar una herramienta que permita al auditor tener una lista de todos los casos que tiene pendientes.

- Opción 3. Separar los servicios cortos de los largos. Los servicios cortos serían atendidos por auxiliares en 3 ó 4 taquillas exclusivas para ellos. Los servicios largos serían atendidos por auditores en 2 ó 3 taquillas exclusivas. Esta opción, al igual que la número 1, incrementa el gasto administrativo, pero en menor proporción.

- Opción 4. Separar, desde la generación de las fichas, los servicios cortos de los largos. Las fichas para servicios cortos tendrán un determinado número de taquillas asignadas para ser atendidas por auxiliares. Los usuarios de estas taquillas esperarán muy poco tiempo, puesto que sus servicios son más ágiles. Los servicios largos, tendrán igualmente unas taquillas asignadas, atendidas también por auxiliares acompañados por el auditor médico. Estos usuarios seguirán con tiempos de espera altos, pero no será un problema, porque sus servicios así lo requieren. Lo importante es que sólo espere quien realmente debe esperar.

Las estrategias 3 y 4 afectan el modelo estandarizado nacional de atención en la sala, el cual se caracteriza por su integralidad y se soporta en que el usuario tenga no más un contacto con empleados de la empresa. Sin embargo, la estrategia 4 podría solucionar los tiempos de espera y mejorar la satisfacción de los usuarios sin incrementar los gastos administrativos. Este modelo supone que tres servidores atienden servicios cortos, y los otros tres atienden los largos. Los tiempos para los servicios cortos se disminuyen por no requerir la intervención del auditor. Los resultados de esta estrategia se muestran en la tabla 9.

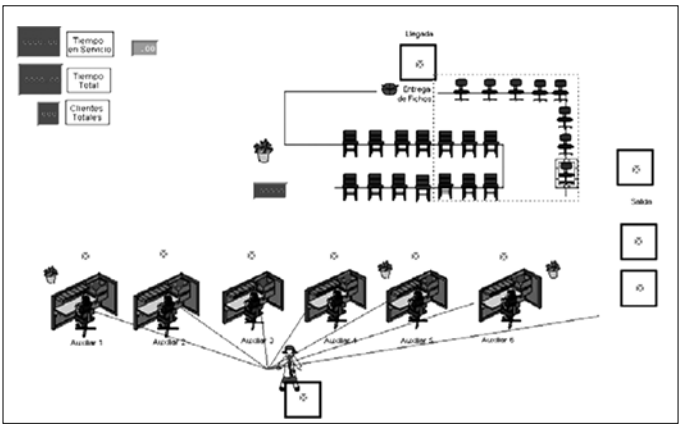

Figura 13. Modelo con fichas diferentes para servicios cortos y servicios largos

Tabla 9. Resultados simulación separación de servicios cortos y largos

		Actual	Cortos	Largos
Número de replicaciones		100	550	210
Nivel de confianza		95%	95%	95%
Error admisible		7%	5%	5%
Parámetros		Media	Media	Media
Número de clientes promedio en el sistema	L	22,01	6,55	12,16
Tiempo promedio de los clientes que esperan en el sistema	W	31,86	11,99	67,69
Tasa de utilización 1	R1	87,33%	77,22%	77,33%
Tasa de utilización 2	R2	85,59%	72,43%	72,63%
Tasa de utilización 3	R3	84,09%	66,33%	66,49%
Tasa de utilización 4	R4	82,62%	84,65%	84,91%
Tasa de utilización 5	R5	80,11%	83,04%	83,33%
Tasa de utilización 6	R6	79,11%	81,31%	81,68%
Número promedio de clientes en cola	Lq	16,01	4,19	9,63
Tiempo promedio de los clientes que esperan en cola	Wq	23,50	7,60	53,48
Tiempo promedio de servicio	μ	8,15	4,32	17,71

Se puede observar que el tiempo de espera promedio en cola (Wq) cae de modo drástico en los servicios cortos, mientras que el de los servicios largos, supera enormemente la meta esperada por la empresa.

Conclusiones

- Las llegadas de clientes al sistema de servicio no se rigen por ningún patrón de demanda especial, lo que garantiza que, con la selección del número adecuado de servidores y su programación organizada, se

puede atender a todos los usuarios dentro de los tiempos de espera que se tienen como meta.

- Al intentar resolver con modelos matemáticos el problema de cuántos servidores requiere el sistema para atender satisfactoriamente a sus usuarios, se encontró que los tiempos entre llegadas de los clientes tenían un comportamiento exponencial, pero que los tiempos de servicio de los auxiliares de la sala no tenía el mismo comportamiento. Y aunque el modelo matemático permitió evidenciar algunas situaciones críticas dentro de la jornada laboral, la simulación discreta permitió una mayor veracidad y validez en los resultados hallados.
- Mediante el análisis matemático y la simulación se encontró que durante la mañana el sistema no tenía por qué generar colas de clientes, que comenzaban a aparecer cuando los servidores se retiraban a almorzar y el sistema perdía capacidad, mientras los clientes seguían llegando. Una vez que el sistema retornaba a su capacidad máxima, la cola que se había formado era tan alta que no se lograba disminuir hasta el final del turno.
- Para atenuar ese problema se evaluaron diferentes propuestas, y las mejores fueron: 1) traslapar y disminuir el receso de almuerzo de cada servidor y 2) tener un auxiliar que trabaje medio tiempo y reemplace a los auxiliares que están almorzando.
- Una mejora importante dentro del cambio en los procedimientos de servicio sería modificar el concepto actual de la sala en la cual todos los auxiliares realizan las mismas funciones y están en capacidad de atender cualquier tipo de problema, garantizando que en cada encuentro de servicio el cliente sólo tenga un contacto con un empleado de la empresa. Esta política obliga a que todos los clientes tengan que esperar mucho tiempo, así requieran un servicio muy corto. La propuesta sería separar aquellos servicios que se sabe que son más largos, porque necesitan un estudio especial para ser autorizados, de aquellos que son muy cortos por su poca complejidad. Esto permitiría que sólo tuvieran que esperar aquellos clientes que tienen casos complejos.
- Las empresas de servicio similares a la estudiada en este proyecto deben realizar con frecuencia estudios que le permitan determinar cuándo el sistema de servicio se está saliendo de control, debido a un aumento inesperado en el número de afiliados, a un incremento en alguna situación concreta o a la creación de un nuevo trámite legal, para garantizar siempre un servicio de calidad.

- El entrenamiento de los auxiliares de la sala y la evolución de las curvas de aprendizaje deben tenerse en cuenta al momento de asignar personas a la sala. Una persona que no esté debidamente formada o que no cumpla con el perfil definido para el cargo puede incrementar el tiempo de espera de los usuarios, por disminuir la capacidad del sistema de servicio.
- Definitivamente, se debe reevaluar el concepto de la sala, porque la integración de todos los servicios en un solo auxiliar contribuye al aumento de los tiempos de espera de todos los usuarios y, por ende, a su insatisfacción.

Bibliografía

CHASE, Richard, AQUILANO, Nicholas y JACOBS, Robert. Administración de producción y operaciones: manufactura y servicios. 8 ed. Santafé de Bogotá: McGraw-Hill, 2000. 885 p.

DAVIS, Mark, AQUILANO, Nicholas y CHASE, Richard. Fundamentos de dirección de operaciones. 3 ed. España: McGraw-Hill, 1999. 598 p.

EPPEN, G. D.; GOULD, F. J.; MOORE J. H.; SCHMIDT, C. P. y WEATHERFORD, L. R. Investigación de operaciones en la ciencia administrativa. México: Pearson Prentice Hall, 2000. 822 p.

HARREL, Charles; GHOSH, Biman, BOWDEN y Royce. Simulating using Promodel. 3 ed. New York: McGraw-Hill, 2000. 603 p.

PULGARÍN, Bernardo. Simulación empresarial (enfoque experimental). 3 ed. Medellín: EIA, 2003. 239 p.

RUSSEL, Roberta y TAYLOR, Bernard III. Operations management. 3 ed. New York: Prentice Hall, 2000. 868 p.

SIPPER, Daniel y BULFIN, Robert. Planeación y control de la producción. 1 ed. México: McGraw-Hill, 1998. 657 p.

VAN DIJK, Nico M. Why queuing never vanishes. En: European Journal of Operational Research. Volume 99, issue 2, 1 June 1997, pp 463-476.

WACKERLY, Dennis; MENDENHALL, William y SCHEAFFER, Richard. Estadística matemática con aplicaciones. 6 ed. México: Thompson, 2002. 853 p.

www.itsom.mex/dii/elagarda Agosto 2004.

www.investigacion-operaciones.com Agosto 2004.

www.scienceofbetter.org Agosto 2004.

ANEXO 1

Análisis de estacionalidad semanal

DICIEMBRE						
Semana	1	2	3	4	5	Totales
Turnos generados	2.073	1.629	2.256	1.742	1.577	9.277
% participación	22,35%	17,56%	24,32%	18,78%	17,00%	100,00%
ENERO						
Semana	1	2	3	4	Totales	
Turnos generados	2.762	1.562	2.066	2.117	8.507	
% participación	32,47%	18,36%	24,29%	24,89%	100,00%	
FEBRERO						
Semana	1	2	3	4	Totales	
Turnos generados	1.435	1.874	1.736	1.668	6.713	
% participación	21,38%	27,92%	25,86%	24,85%	100,00%	
MARZO						
Semana	1	2	3	4	Totales	
Turnos generados	1.799	1.727	1.910	1.670	7.106	
% participación	25,32%	24,30%	26,88%	23,50%	100,00%	
ABRIL						
Semana	1	2	3	4	Totales	
Turnos generados	2.170	1.795	1.938	1.711	7.614	
% participación	28,50%	23,57%	25,45%	22,47%	100,00%	
MAYO						
Semana	1	2	3	4	Totales	
Turnos generados	1.943	1.643	1.845	1.737	7.168	
% participación	27,11%	22,92%	25,74%	24,23%	100,00%	
JUNIO						
Semana	1	2	3	4	5	Totales
Turnos generados	1.190	1.166	1.642	1.558	1.660	7.216
% participación	16,49%	16,16%	22,75%	21,59%	23,00%	100,00%
JULIO						
Semana	1	2	3	4	Totales	
Turnos generados	1.460	1.678	1.504	1.664	6.306	
% participación	23,15%	26,61%	23,85%	26,39%	100,00%	
AGOSTO						
Semana	1	2	3	4	Totales	
Turnos generados	1.703	1.565	1.446	1.726	6.440	
% participación	26,44%	24,30%	22,45%	26,80%	100,00%	

ANEXO 2

Diferencia de medias propuesta 1, alternativas 1 y 2

Si queremos probar $H_0 : \mu_1 - \mu_2 = D_0$, frente a la hipótesis alternativa, el estadístico de la prueba está representado por:

$$Z = \frac{(\overline{Y}_1 - \overline{Y}_2) - D_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Donde σ_1^2, σ_2^2 son las varianzas poblacionales. En el caso de muestras grandes, n > 30, las varianzas muestrales proporcionan buenas estimaciones de sus correspondientes varianzas de población. Si α = 0,05, se rechaza H_0 para $|z| > z_{\alpha/2} = z_{0.025} = 1.96$

	Actual		Alternativa 1		Alternativa 2	
Número de replicaciones	100		200		200	
Parámetros	Media	Desv Est	Media	Desv Est	Media	Desv. Est.
Tiempo promedio de los clientes que esperan en cola Wq	23,50	9,57	21,91	8,79	17,38	8,81

- Para comparar la situación actual con la alternativa 1, traslapar los almuerzos se tiene: Z = 1,39, por lo tanto, las medias son iguales.
- Para comparar la situación actual con la alternativa 2, traslapar y disminuir los tiempos de almuerzo, se tiene: Z = 5,36, por ello, las medias son diferentes.

Diferencia de medias propuesta 2, alternativas 1 y 2

	Actual		Alternativa 1		Alternativa 2	
Número de replicaciones	100		500		500	
Parámetros	Media	Desv Est	Media	Desv Est	Media	Desv Est
Tiempo promedio de los clientes que esperan en cola Wq	23,50	9,57	10,02	6,68	13,63	8,60

- Para comparar la situación actual con la alternativa 1, utilizar 7 servidores todo el turno se tiene: Z = 13,46, por consiguiente, las medias son diferentes.
- Para comparar la situación actual con la alternativa 2, utilizar un séptimo servidor sólo durante la hora del almuerzo, se tiene: Z = 9,56, por tanto, las medias son diferentes.