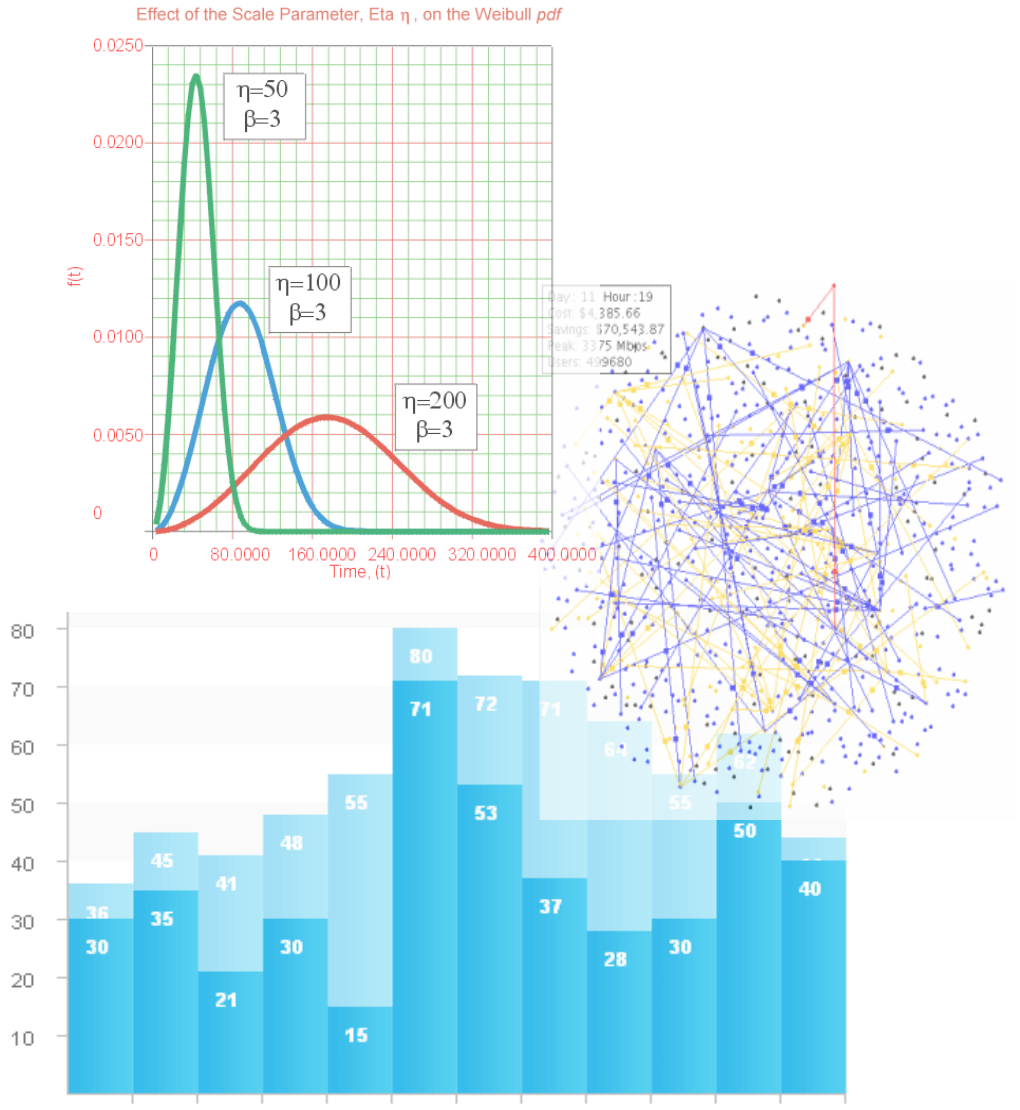
Práctico Especial II

Análisis estadístico de datos simulados

A. Perez Paladini, C. E. Budde
FaMAF – UNC – Argentina
Modelos y simulación
22 de Junio del año 2010



INTRODUCCIÓN

El objetivo de este trabajo fue hacer uso de las diferentes herramientas de estimación estadística para plantear una hipótesis acerca de la densidad de probabiliad teórica a la que pertenecía un conjunto de datos muestrales, los que fueron generados mediante simulaciones.

El sistema modelado representaba un servidor con cola de espera, donde el tiempo medio de arribo de cada cliente era de ¼ de hora obedeciendo una distribución exponencial, y el tiempo que le dedica el servidor a cada uno también estaba exponencialmente distribuido con tasa igual a 4,5 hora^{-1}. El sistema recepta cliente sólo 8 hs al día, y puede haber a lo sumo 4 clientes en cola de espera en cualquier instante de tiempo dado.

La información relevante que quiso obtenerse de la simulación era el tiempo medio de espera de los clientes en el sistema[1]. Se consideró que los tiempos de espera entre días disjuntos corresponden a v.a. Independientes. Se propuso como estimador de dicha magnitud para el i-ésimo día a la relación:

$$\hat{e}_i = \frac{\bar{D}_i}{\bar{N}_i}$$

donde: $\bar{D}_i$ = suma durante el día de los tiempos de residencia de los clientes en el sistema
$\bar{N}_i$ = # total de clientes atendidos en el i-ésimo día

Se simularon 500 días de funcionamiento del sistema, y la estimación final del tiempo medio de espera se calculó como el promedio de los $\hat{e}_i$ para los 500 días.

SIMULACIÓN DEL SISTEMA

Para modelar el funcionamiento del servidor se implementaron únicamente dos módulos. Uno correspondió al manejo de la cola de espera (módulo *queue*), y el otro a la operación del sistema en sí (módulo *simulation*). De esta forma se logró un balance entre la hipermodularización y el hardcoding de todas las rutinas en un único main.

Sólo para ser tediosos se presentan a continuación los 500 datos obtenidos:

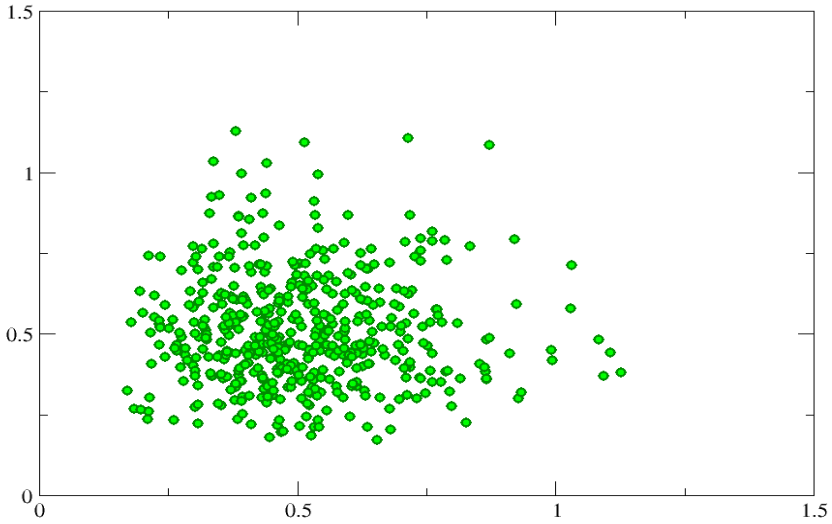
1 Espera total: desde el ingreso del cliente al sistema hasta la finalización de su atención por parte del servidor

Día	+0	+100	+200	+300	+400
1	0.69691283	0.34951643	0.39088294	0.39462817	0.40777028
2	0.46591294	0.73878001	0.35325642	0.30390100	0.70930190
3	0.61324717	0.79552089	0.27672775	0.42613300	0.78540039
4	0.62713977	0.32164969	0.69706853	0.44297979	0.78949078
5	0.43487959	0.54365346	0.29782397	0.33804220	0.72772129
6	0.79845274	0.44664823	0.57991162	0.58027040	0.73838817
7	0.27463680	0.57698765	0.44801640	0.49178384	0.72532646
8	0.39581805	0.31774544	0.42292684	0.72186251	0.56325733
9	0.61482508	0.65945600	0.56763656	0.63431118	0.35931898
10	0.33745189	0.30188954	0.53018329	0.70200908	0.52568936
11	1.03244970	0.38647698	0.35694668	0.52541484	0.36773989
12	0.71113341	0.38585900	0.36212668	0.31519737	0.53190886
13	0.53198103	0.86455440	0.43313577	0.76255409	0.44075602
14	0.21226596	0.38364005	0.44231241	0.81515202	0.28666316
15	0.25868816	0.60494823	0.46769171	0.36105238	0.41836837
16	0.54441951	0.44136049	0.39872961	0.51085046	0.46244994
17	0.48150597	1.02910055	0.40371163	0.46225435	0.37874977
18	0.33457725	0.57706946	0.59254973	0.38042411	0.29324107
19	0.66960955	0.57143922	0.53861496	0.43985177	0.63163819
20	0.43447117	0.43141617	0.68037763	0.53195865	0.30727748
21	0.32797226	0.62871161	0.20172344	0.59332470	0.28042842
22	0.42882973	0.45533540	0.56471891	0.62350785	0.35363952
23	0.64376620	0.52165762	0.54497224	0.71186266	0.37563247
24	0.76152612	0.59140248	0.52389854	0.36406507	0.60616309
25	0.78811639	0.50501984	0.27685799	0.36709085	0.63542766
26	0.38187299	0.71604715	0.48858092	0.52410181	0.60257047
27	1.12633539	0.62257731	0.38743079	0.45073767	0.24419866
28	0.37992633	0.46694376	0.51395468	0.47083010	0.42891879
29	0.70445472	0.41874765	1.09284389	0.39280982	0.53724983
30	0.44796941	0.77356844	0.36792478	0.81063263	0.76232240
31	0.64053035	0.55604719	0.73769374	0.53315510	0.43826238
32	0.54113718	0.55782150	0.36375172	0.38108906	0.51083818
33	0.99192814	0.46585580	0.44607587	0.34607936	0.64032995
34	0.45073163	0.48960725	0.17795885	0.42881613	0.47699556
35	0.49733937	0.65825735	0.53493972	0.71512782	0.44957161
36	0.39899055	0.63998329	0.56858043	0.62247922	0.30807312
37	0.45011453	0.40639640	0.62330857	0.74875745	0.33839093
38	0.42604915	0.43750991	0.46506968	0.31670868	0.70672070
39	0.71482579	0.57111600	0.60110197	0.52533853	0.39250676
40	1.10613205	0.76191542	0.43209979	0.74561163	0.29168921
41	0.44271592	0.34988862	0.37119384	0.46992081	0.59003803
42	0.34741502	0.92881430	0.32567379	0.56761791	0.59249182
43	0.61492012	0.29894784	0.49865954	0.46315784	0.48047396
44	0.43881893	0.71923760	0.38880751	0.53913799	0.39112153
45	0.56120647	0.86638959	0.39349862	0.30829486	0.55890880
46	0.67979944	0.35740190	0.55066192	0.69767638	0.50865656
47	0.71926536	0.52163253	0.75804368	0.58216209	0.52253093
48	0.39635455	0.65434967	0.53412475	0.38139439	0.66421879
49	0.77297954	0.17165992	0.49264720	0.53409987	0.52117654

Día	+0	+100	+200	+300	+400
50	0.55548448	0.32321150	0.40987450	0.33018396	0.31730246
51	0.50797721	0.51057733	0.43965010	0.87173324	0.45229574
52	0.29881312	0.61291857	0.93485146	1.08451229	0.30990363
53	0.40059054	0.46340850	0.31732711	0.48243220	0.59853426
54	0.60567495	0.59352221	0.62690382	0.46433295	0.68851637
55	0.33647471	0.51765088	0.60465197	0.45637184	0.59038399
56	0.36893412	0.24384048	0.68188450	0.48871344	0.78030788
57	0.56542859	0.58910875	0.52064710	0.50219532	0.53447042
58	0.44467994	0.43447234	0.28311034	0.37173792	0.86747270
59	0.56279032	0.62112908	0.43432166	0.61533764	0.36088296
60	0.43796582	0.55323456	0.36501047	0.39207211	0.62200563
61	0.69135305	0.73086316	0.44444543	0.99493118	0.60727477
62	0.63867124	0.30074143	0.63464643	0.41771719	0.34492858
63	0.70004117	0.53607469	0.55965398	0.48857327	0.70778413
64	0.50511609	0.50779819	0.51475887	0.39996093	0.63223697
65	0.64984518	0.35354306	0.67982987	0.40489045	0.64393502
66	0.46974466	0.62602239	0.26620999	0.46369593	0.44397334
67	0.19536420	0.39906375	0.44793770	0.37619865	0.52679404
68	0.63101817	0.40635310	0.49117259	0.55699997	0.18330294
69	0.32404755	0.85281298	0.54533920	0.63773306	0.26617325
70	0.48198726	0.40616831	0.31946154	0.40694500	0.46436799
71	0.59785534	0.41053868	0.52222323	0.30740391	0.44253558
72	0.86615378	0.69108915	0.43257334	0.22302164	0.30492097
73	0.48259678	0.41730382	0.87221595	0.61802277	0.73893916
74	0.61503244	0.38089180	0.48750074	0.53785062	0.36182730
75	0.34964356	0.33466701	0.33409195	0.54059898	0.55137116
76	0.43393430	0.36984942	0.92471944	0.82706721	0.66549216
77	0.35807675	0.44056099	0.59196218	0.22384411	0.45318188
78	0.42058440	0.71007786	0.53731538	0.55171567	0.49653435
79	0.43390298	0.39579983	0.46165873	0.55795437	0.56020677
80	0.46743259	0.55394324	0.48087358	0.26182281	0.50952458
81	0.29801852	0.49565926	0.40586551	0.23228533	0.39471347
82	0.76996836	0.71254813	0.43222607	0.54144671	0.60538984
83	0.57537658	0.31019600	0.30199151	0.21000838	0.42751536
84	0.66018896	0.48869489	0.27278399	0.23508141	0.48927331
85	0.59600082	0.64562020	0.50283480	0.73853045	0.35389252
86	0.41040460	0.50471540	0.54103225	0.75870963	0.59043316
87	0.92128294	0.21431899	0.23305399	0.38428120	0.45428021
88	0.79169871	0.30243315	0.46604743	0.23393824	0.32251566
89	0.38628186	0.36870286	0.23289314	0.51972791	0.48659894
90	0.37060661	0.49591348	0.52883954	0.63620287	0.47273558
91	0.75153666	0.54609290	0.33300827	0.21229661	0.19651320
92	0.46195894	0.36436424	0.37812374	0.74248835	0.26343406
93	0.21571337	0.52948242	0.64794804	0.52128766	0.45429237
94	0.50271616	0.64848145	0.58887958	0.42388836	0.45207546
95	0.46511656	0.71527257	0.30203647	0.39372328	0.34744080
96	0.83443622	0.48307976	0.49982045	0.25202803	0.28329822
97	0.77082047	0.39142271	0.68294650	0.51679764	0.45573860
98	0.57610924	0.35379216	0.54983244	0.71786280	0.53225941
99	0.33756749	0.47447502	0.38715127	0.41087618	0.91114277
100	0.77779793	0.46248339	0.86200408	0.21795029	0.43753273

ACTIVIDAD 1: INDEPENDENCIA ESTADÍSTICA DE LA MUESTRA

Analizando los datos muestrales se obtuvo el siguiente gráfico de dispersión:



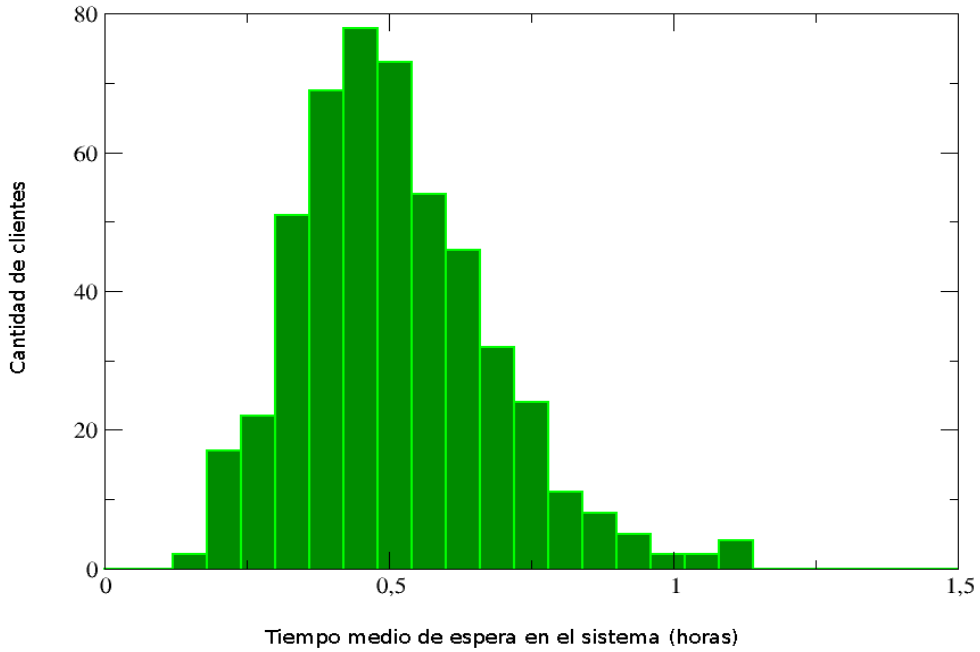
Analizándolo a simple vista resulta evidente que no existen indicios de correlación alguna entre los datos generados a lo largo de los días sucesivos. Se sigue de esto que la hipótesis de independencia entre los estimadores $\hat{e}_i$ de cada día fue satisfecha, y por consiguiente la muestra obtenida es apta para ser objeto de un análisis estadístico insesgado.

ACTIVIDAD 2: ESTADÍSTICAS DE LA MUESTRA

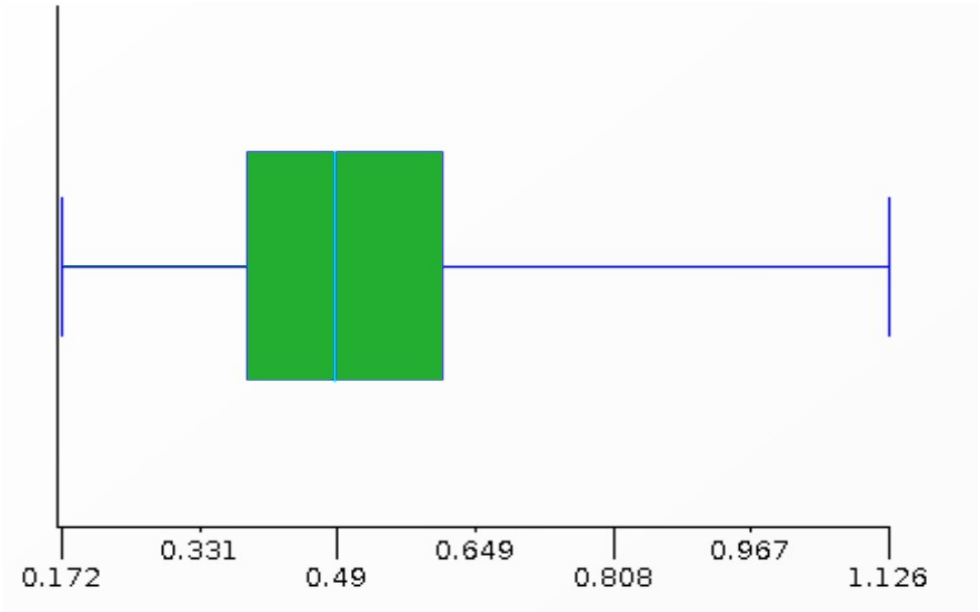
a) Magnitudes significativas

Media	0.50934599
Varianza	0.02919431
Extremos	
Minimo	0.17165992
Maximo	1.12633539
Cuartiles	
Q1	0.38729103
Q2 (Mediana)	0.48863790
Q3	0.61308287
Skewness	0.70160439

b) Histograma de los datos



b) Box-plot



Tiempo medio de espera en el sistema (horas)

Los whiskers izquierdo y derecho representan los valores máximo y mínimo observados respectivamente.

ACTIVIDAD 3: ESTIMACIÓN DE PARÁMETROS

En base a los datos muestrales se propusieron 3 distribuciones de probabilidad posibles para la población de donde fueron obtenidos. Como naturalmente se desconocían los parámetros de las funciones densidad de dichas distribuciones, éstos tuvieron que ser estimados mediante la misma muestra. Las distribuciones propuestas y sus parámetros estimados son:

a) Gamma (α,β)

Función densidad: $\mathrm{f}(x)=\frac{\beta^{\alpha}}{\Gamma(\alpha)}x^{\alpha-1}\mathrm{e}^{-\beta x}$

EMV: se emplearon las fórmulas aproximadas

$T=[\ln \bar{X}(n)-\sum_{i=1}^{n}\ln\frac{X_i}{n}]^{-1}$ $\quad\hat{\alpha}=\mathrm{f}(T)$ * $\quad\hat{\beta}=\frac{\bar{\mathrm{X}}\,n}{\hat{\alpha}}$

Estimadores calculados:

$\hat{\alpha}=8.95$ $\quad\hat{\beta}=0.0569102$

b) Normal (μ,σ)

Función densidad: $\mathrm{f}(x)=\frac{1}{\sqrt{2\pi\sigma^2}}\mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}$

EMV:

$\hat{\mu}=\bar{\mathrm{X}}(n)$ $\quad\hat{\sigma}=\sqrt{\frac{n-1}{n}S^2(n)}$

Estimadores calculados:

$\hat{\mu}=0.50934599$ $\quad\hat{\sigma}=0.17069248$

c) Lognormal (μ,σ)

Función densidad: $\mathrm{f}(x)=\frac{1}{x\,\sigma\sqrt{2\pi}}\mathrm{e}^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}$

* donde f(T) está dada de forma numérica en una tabla de la bibliografía.

EMV:

$$\hat{\mu} = \frac{\sum_{i=1}^{n} \ln X_i}{n} \qquad \hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n} (\ln X_i - \hat{\mu})^2}{n}}$$

Estimadores calculados:

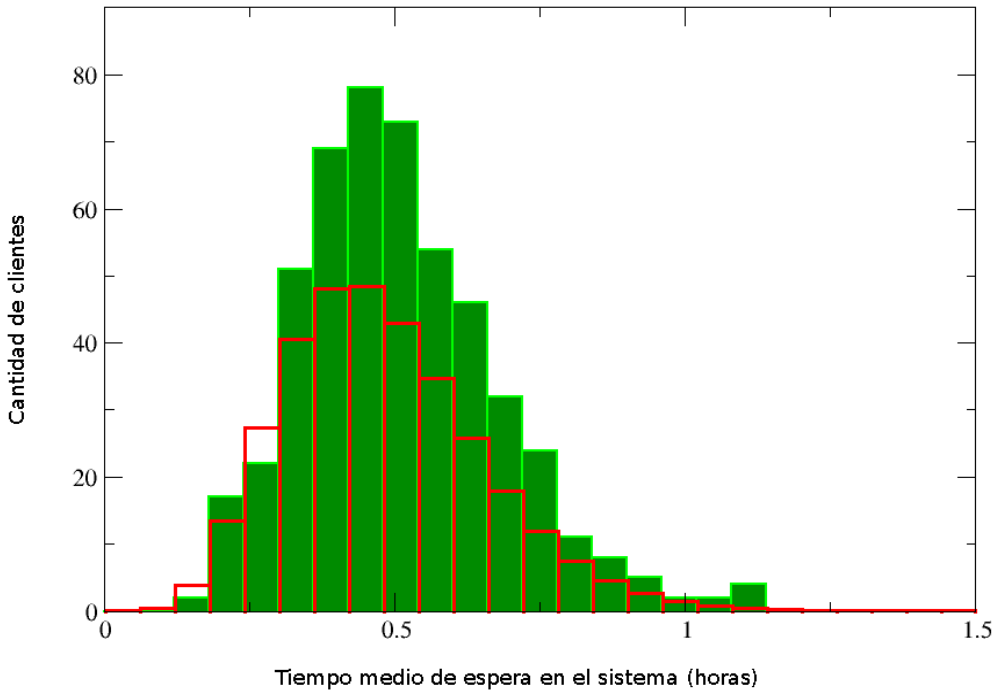
$$\hat{\mu} = -0.73124930 \qquad \hat{\sigma} = 0.34184440$$

ACTIVIDAD 4: PRUEBA DE HIPÓTESIS

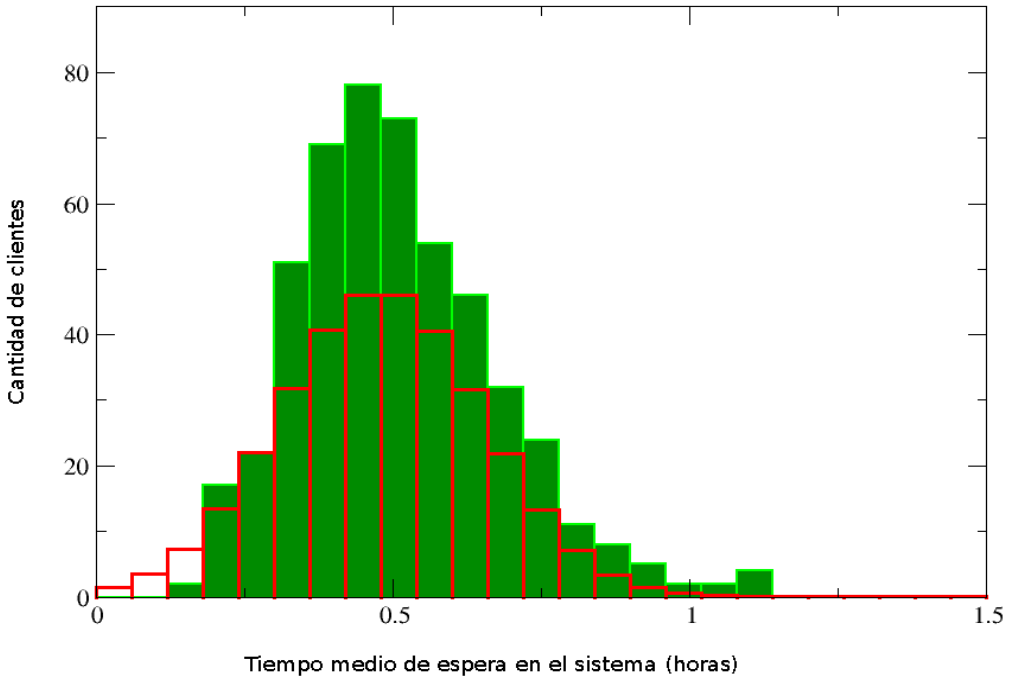
Para cada una de las 3 distribuciones teóricas inferidas se comparó visulamente el histograma generado con los datos, con el correspondiente a la función densidad de la distribución respectiva.

En los siguientes gráficos de barras las columnas verdes rellenas se corresponden con el histograma muestral ya presentado, y las barras rojas superpuestas modelan el histograma de las distribuciones de probabilidad teóricas.

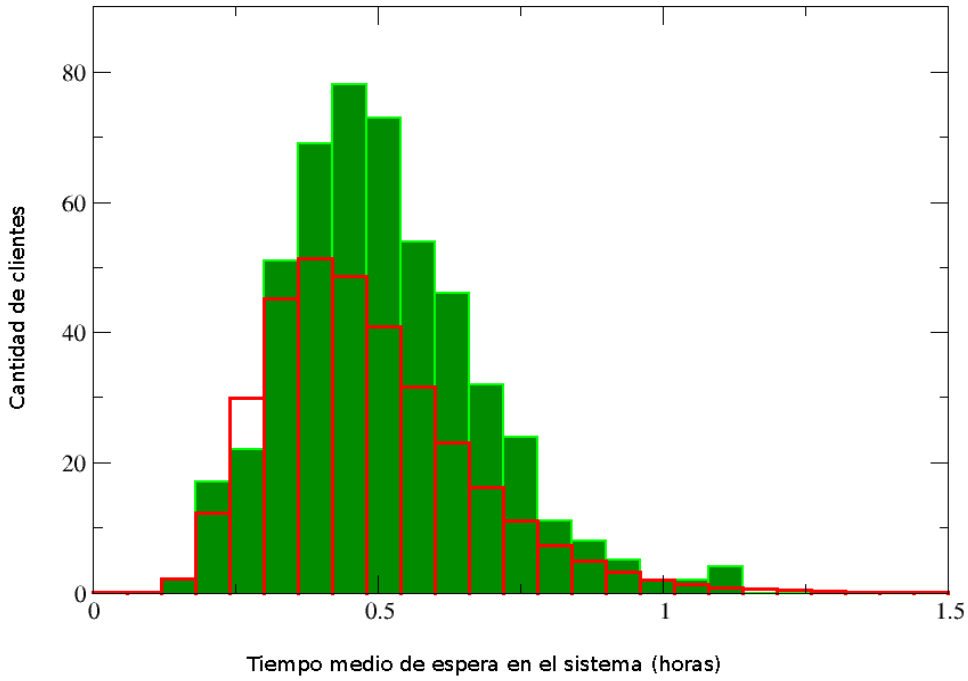
a) Gamma (α,β)



b) Normal (μ,σ)



c) Lognormal



Como la comparación cualitativa visual no es razón suficiente para decidirse por una de las tres opciones, también se realizaron pruebas cuantitativas. En particular se evaluaron los estadísticos de Ji-cuadrado y de Kolmogorov-Smirnov para estimar el p-valor de las muestras con respecto a cada una de las hipótesis. Los resultados fueron:

a) Gamma (α,β)

Ji-² (aprox)	Ji-² (sim)	K-S
T = 18.17995059 **p-valor** = 0.69526750	**T** = 18.17995059 **p-valor** = 0.47710000	**D** = 0.02135111 **p-valor** = 0.97420000

b) Normal (μ,σ)

Ji-² (aprox)	Ji-² (sim)	K-S
T = 127.93357767 **p-valor** = 0.00000000	**T** = 127.93357767 **p-valor** = 0.16420000	**D** = 0.06239674 **p-valor** = 0.03580000

c) Lognormal (μ,σ)

Ji-2 (aprox)	Ji-2 (sim)	K-S
T = 21.07446884 **p-valor** = 0.51614022	**T** = 18.17995059 **p-valor** = 0.46610000	**D** = 0.03443693 **p-valor** = 0.58140000

Consideraciones sobre el cálculo del p-valor

Simulaciones

Como no se disponía a priori de un valor para los parámetros de las distribuciones hipotéticas contra las que se compara la muestra, éstos tuvieron que ser estimados en base a los datos empíricos. El marco teórico disponible indica entonces que ante la necesidad de simular nuevas muestras para ajustar el p-valor, cada nueva simulación debe ser utilizada para recalcular los estimadores para los parámetros. Así, el nuevo cálculo del valor-p emplea en cada simulación la distribución $F_{\hat{\sigma}}$ que surge de reestimar los parámetros desconocidos mediante los valores simulados.

De haber seguido este camino habría sido necesario disponer de un método automatizado que ante un conjunto de valores muestrales haga una estimación de los parámetros de la función de distribución de que se trate. Pero el trabajo no acabaría llí, pues en función de estos nuevos parámetros estimados la rutina debería construir una nueva función de distribución $F_{\hat{\sigma}}$ con la cual se calcularía el estadístico correspondiente.

Proceder de esta forma habría sido altamente dificultoso, por no decir prácticamente imposible, a causa de las complicaciones que plantea la estimación de ciertos parámetros de las funciones densidad, como es α en el caso de la distribución Gamma. Además tal proceder habría hecho que el tiempo de cálculo se prolongue de manera exponencial, quizás quedando fuera del alcance de ordenadores de baja gama como los que se disponen para el proyecto.

Por todo esto se decidió tomar un curso de acción diferente. Se supusieron teóricamente plausibles las hipótesis de que la muestra provenía de cada una de las tres distribuciones de probabilidad, **asumiendo que los parámetros que regulan su comportamiento no fueron estimados en base a la muestra**, sino que fueron depositadas por intervención divina en los rincones más iluminados de nuestras mentes.

De esta forma las complicaciones se simplifican enormemente .Esto gracias a que, de ser necesarias, ante cada nueva simulación puede emplearse aquella función de distribución de probabilidad calculada (una única vez) con la muestra original para evaluar los estadísticos derivados de las simulaciones.

Funciones de distribución acumulada para χ^2 y K-S

A la hora de calcular el p-valor de la muestra empleando el método de Ji-2 se disponía de dos caminos a seguir. Una opción era hacer uso del resultado teórico que asegura que el estadístico T tiene distribución aproximadamente Ji-cuadrada con k-1-m grados de libertad, donde **k** es el # de intervalos en el que se dividió la muestra (k=25 en los tres casos) y **m** es el # de parámetros estimados de la distribución teórica (m=2 en los tres casos). Este camino sugiere que el valor-p = P(T>***t***), donde ***t*** es el estadístico de la muestra y T es una v.a. ~ Ji-2 con k-1-m grados de libertad.

La otra opción consistía en ajustar este valor mediante simulaciones. El método supone generar muestras según la distribución teórica hipotética (cuyos parámetros fueron estimados con la muestra original), reestimar los parámetros de la distrib. con estos nuevos datos simulados, y finalmente calcular en base a esos datos y a esa distribución un estadístico *Ti* correspondiente a la i-ésima simulación. Luego: $valor - p = \frac{\#\{i : Ti > t\}}{\#\{simulaciones\}}$

Como no se disponía a priori de un nivel de rechazo para decidir si la hipótesis planteada debía ser o no rechazada, no era claro cual de las dos opciones era la más indicada para este proyecto. Luego de deliberar largo y tendido sobre el asunto se había decidido emplear la 1º estrategia. Haciendo notar que la reestimación de los parámetros tras cada simulación quedó descartada por razones expuestas anteriormente, la razón principal de tal decisión se basó en la forma que tiene este test de calular el estadístico de una muestra.

Para evaluar el *Ti* de una simulación es necesario disponer de los valores que indican la probabilidad "acumulada" de las distribuciones teóricas en los intervalos de división del rango. Las funciones Normal, Gamma y Lognormal son conocidas por lo horripilante de sus funciones densidad. Tan es así que se desconocen formas cerradas para

sus funciones de distribución acumulada. Todo esto llevó a los autores a pensar en un comienzo que sería vano todo intento por tratar de emplear Ji-2 con simulaciones.

Sin embargo demostrando una increíblemente baja estima por la vida social y la salud general, ellos se empeñaron en tratar de aproximar estos valores para los intervalos mediante cualquier medio a su alcance. Un primer intento constituyó el uso de Montecarlo para aproximar el área por debajo de la curva de cada función densidad dentro de los límites impuestos por cada intervalo.

Este primer intento fracasó rotundamente, (se cree) debido a que el # de simulaciones Montecarlo no podía hacerse lo suficientemente grande como para estimar de manera fehaciente esta integral definida. Con él se obtuvieron datos ridículamente bajos para los valores-p, que incluso no guardaban relación con los resultados cualitativos presentados por la superposición de histogramas.

Por ejemplo llegó a aparecer el valor-p de la muestra en relación con la distribución gamma muy por debajo del de la distribución lognormal, a mitad de camino con la distribución normal. Como el aumento en la precisión decimal de dichas aproximaciones es muy lento en ralción con el aumento en el # de simulaciones Montecarlo, se abandonaron los intentos por esta vía.

Se decidió entonces hacer uso de herramientas más poderosas. Específicamente se empleó el motor matemático de libre distribución ofrecido por wolfram alpha[2], a quien los autores le estarán eternamente agradecidos. Esta página no sólo brindó formas simplificadas y mucho más digeribles de las funciones densidad dados ciertos valores fijos de los parámetros, sino que además llegó al punto de ofrecer una forma analítica de las funciones acumuladas tan codiciadas. En conjunto con la función estimadora de error "erf" cuyo algoritmo se obtuvo de la página web "Numerical Recipes in C"[3]

Esto también solucionó problemas con el test de Kolmogorov-Smirnov, que si bien estaba generando datos en concordancia relativa con los resultados cualitativos de los histogramas (la distribución normal aparecía como mala aproximación, y las otras dos se peleaban la punta), aún así los valores absolutos obtenidos para p-valor de la muestra eran demasiado bajos, por el orden de $cte*10^{-3}$.

Al cambiar la función de distribución acumulada estimada mediante Montecarlo por la forma analítica ofrecida por wolfram-alpha en el cálculo de los estadísticos K-S, los valores-p adquirieron magnitudes más respetables, que son las que se presentaron como resultados.

2 http://www.wolframalpha.com/

3 http://www.fizyka.umk.pl/nrbook/bookcpdf.html

CONCLUSIÓN

A modo de cierre se señala que como los valores-p de la hipótesis de que la muestra proviene de una distribución **Gamma** son los mayores de entre las tres opciones, se elige a este modelo como el que mejor explica el comportamiento del tiempo medio de espera de los clientes en el sistema. Esto se debe a que una vez fijato un nivel de rechazo α, es menos probable que se rechace la hipótesis de que la muestra proviene de ésta distribución.

Si bien las diferencias de estos indicadores según el test de χ^2 son muy pequeñas entre los casos gamma y lognormal, ese método está pensado para la comparación en situaciones en las que se trabaja con distribuciones discretas. Como el proyecto lidia exclusivamente con funciones de distribución contínuas, se asume que Kolmogorov-Smirnov es el más fiable de los tests, y en su caso el p-valor calculado bajo la hipótesis de que la muestra proviene de una distrib. Gamma supera en casi un 40% al de la Lognormal, inclinando la balanza a favor del 1º como ya se indicó.