Pages 2-5: AB-Testing
Pages 6-11: MIMIC3d

Is WholeCart's Old or New Website Better?

We were commissioned to conduct analysis on behalf of WholeCart LLC, owner of the multi-store WholeCart grocery chain, which claims a large footprint in Ontario. It offers services such as online grocery ordering through a website as well as home delivery and contactless grocery pickup.

You (the board of WholeCart LLC) asked us to investigate the rollout of the redesign of your website used for making online grocery orders. In our analysis, we focused on whether using the new website was more conducive to users signing on to the WholeCart Membership Program (WCMP), which for a flat monthly rate offers savings on select products.

We set up a test of the two websites as follows: When users connected to your website, they were divided into either the “control” or “treatment” group with roughly 50/50 odds. Users in the control group were routed to your old grocery ordering website, as expected, though they were encouraged to use the new website and given a link to the new version. Users in the treatment group were rerouted to the new grocery website, but given the choice to revert to the old website. We then checked whether the user signed up for the membership in that session. Finally, we used the data collected to test whether there was a higher membership signup rate for users who used the new website rather than the old one.

As part of our analysis, we collected the following information:

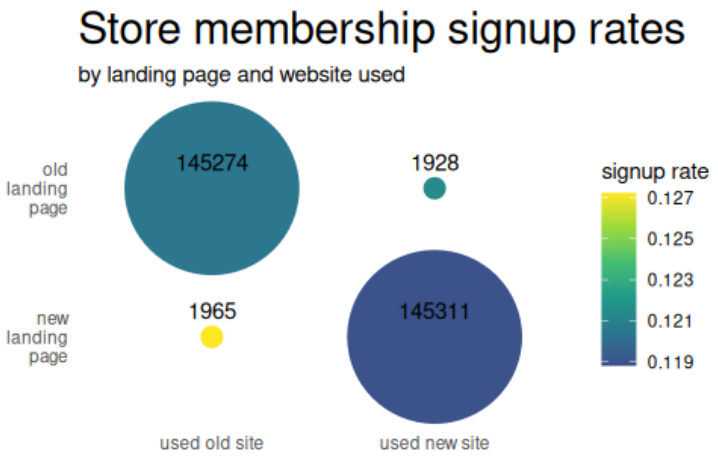
Attribute	**Explanation**	**Values**
user_id	Identifies the user interacting with the website. No user connected more than twice in the time period for which we collected data.	Factor with integer value between “630000” and “945999”.
timestamp	Identifies the time of the interaction between the user and the website. Isn’t used in this analysis.	Timestamp between “2017-01-01 13:42:15” and “2017-01-24 13:41:54”.
group	Defines whether the user was in the control or treatment group.	Factor with value “control” or “treatment”.
landing_page	Defines which page the user interacts with, either the old page or the new page. The old page represents the website that is currently deployed and the new page represents the new website in production, which is being tested.	Factor with value “old_page” or “new_page”.
converted	Whether or not the customer signed up for our store membership program at the time.	Factor with value “0” (not converted) or “1” (converted).

For example:

user_id	timestamp	group	landing_page	converted
791080	2017-01-06 20:53:56	control	old_page	0
791080	2017-01-13 13:31:02	treatment	old_page	0

These two data points indicate two interactions between the same user and our website. The first interaction takes place around 9 p.m. on January 6th. At this time, the user is assigned to the control

group, meaning that they are sent to the old website, but encouraged to use the new one. During this session they do not sign on to our membership program. The second interaction takes place a week later on January 13th at 1:30 p.m. This time, the user is assigned to the treatment group. They are redirected to our new webpage, but given the option of using our old one. This customer chooses to use our old webpage. Again, they do not sign on to our store membership program.

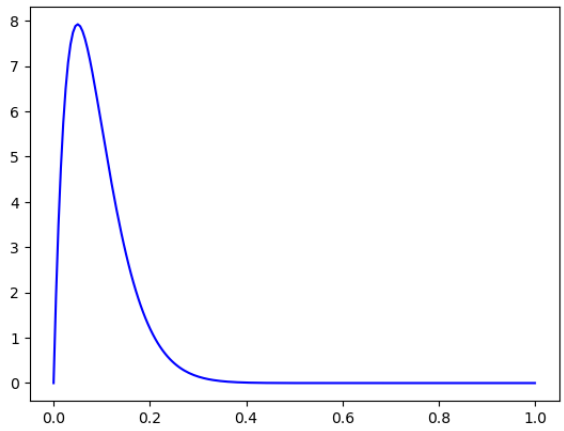


A quick overview of the data indicates that most of our customers who were sent to a given site stayed on that site, rather than electing to use the other website version. What's remarkable about the data is that even between the four categories shown, being in any of the four groups (e.g., "old landing page"/"used new site") didn't change a customer's expected chance of signing up by more than one percentage point (0.01).

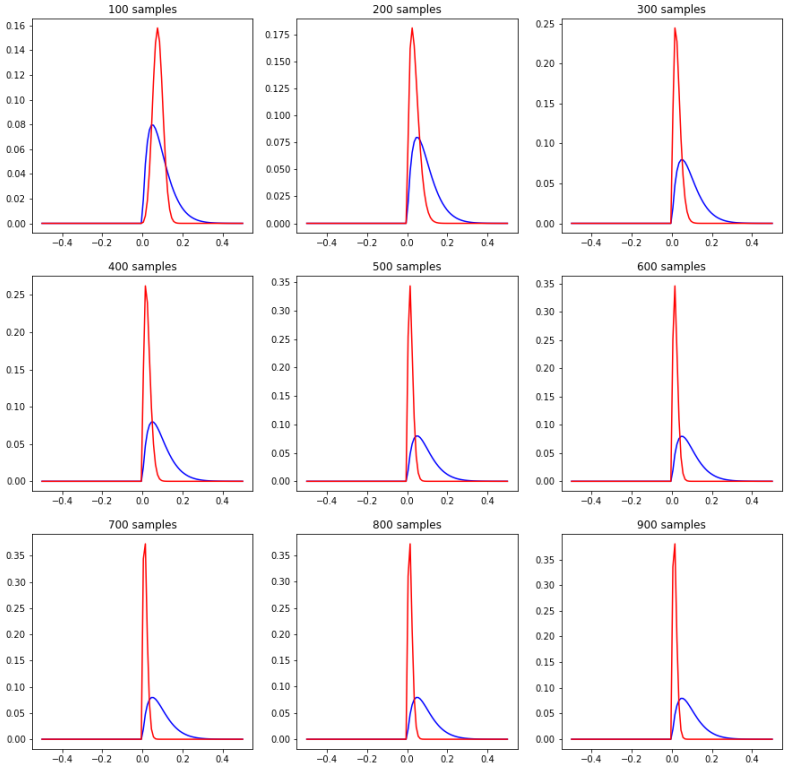
Nevertheless, the small differences in signup rate are worth investigating closely. Notably, those customers who were directed to the new website but instead elected to use the old one were the most likely out of the four groups to sign up for your store membership. From the data above, we also have the increasing suspicion that the new website might in fact be *worse* at getting customers to sign up for the membership plan. After all, those who used the new website had a lower signup rate than those who used the old one.

Given that there is some uncertainty in a given estimation of the rate at which customers on the old or new websites will sign up for the WholeCart Membership Program (called the "conversion rate"), we used methods of Bayesian analysis to determine whether customers sent to the new landing page would have a higher conversion rate than customers sent to the old landing page.

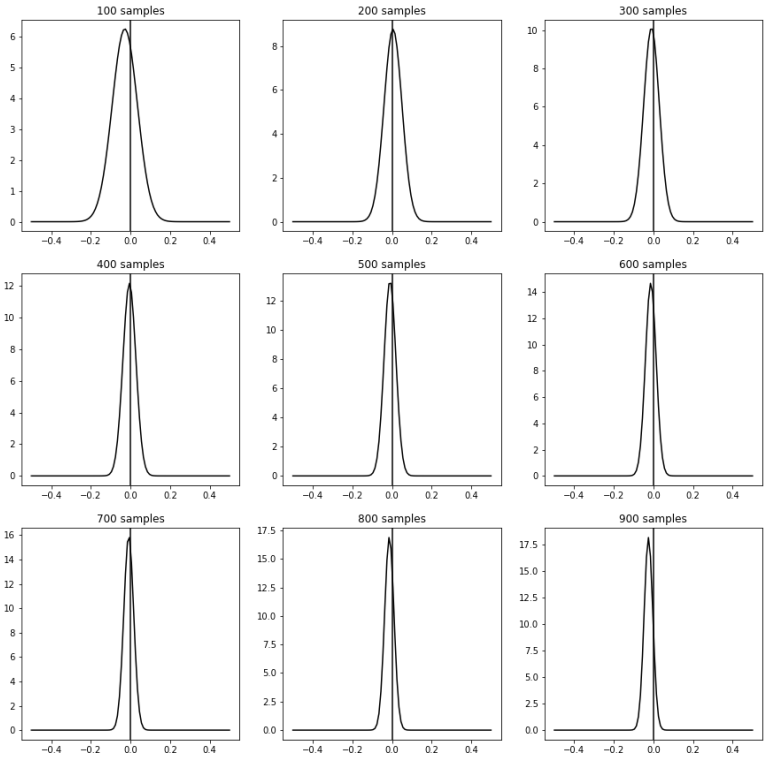
Bayesian methods allow us to quantify any uncertainty about our predictions of the conversion rates. When we look at whether the new website has a better true signup rate than the old one, we can represent our suspicions as a statistical distribution (right). In the technical language of Bayesian analysis, this is our "prior distribution" of **Beta(2, 20)**.The skew of the graph towards zero (with a peak at 1/11) reflects our suspicions that the new website does not in fact have a higher "true" conversion rate than the old one.



In order to determine if the chosen prior distribution was informative and to verify if the new page had a better conversion performance than the current page, we conducted Bayesian inference. The graphs to the right show the distribution of the prior in blue, and the distribution of the posterior in red. After taking about 300 data points (or, iterating three times), the red curve begins looking like the likelihood function itself (see figure), which describes the distribution of the difference between the proportion of conversion of the new page and the old page.



Due to this behavior, we can determine that the prior distribution (Beta(2, 20)) is not an informative prior. Hence we can say that regardless of the original guess ("prior distribution") we chose, at a data size of 300 the priors have become largely irrelevant.



In examining the likelihood distributions, we can see that most of the probability density is located on the negative sides of the graph, meaning that it is most likely that the current page has a greater proportion of conversion than the new page. Also, this also offers a hint to the reason why the prior distribution isn't informative, since the beta distribution is equal to 0 for all negative values, which is where the difference of the sample proportions is most likely located.

We conducted a further test to confirm our suspicions that the true conversion rate for the new website is in fact less than the one for the old one. This was a Monte Carlo simulation in which we compared the performance of the data that we collected to synthetic data with different true conversion values.

The synthetic data was created to have conversion rates similar to our dataset, but with known values that we decided on. For example, we created a set where the proportions are in fact equal, and one of each where one true conversion rate is greater than the other. We administered a test where at each iteration we calculated the suspected chance that the old webpage has a higher true conversion rate than the new release. Then we simply compared the behaviour of our data against the synthetic data.

In the figure below, the green line represents the data collected. The red line represents synthetic data where the proportion of conversion of the current page is greater than the proportion of conversion of the new page. The green and red data have similar outcomes under the test that data sampled from the old website has a higher sample conversion rate than data sampled from the new site. Moreover, the green line diverges from the equal-conversion-rate data (black) and the better-website data (blue). Therefore, we can reject the hypothesis that both websites have the same proportion of conversion for the hypothesis that the proportion of conversion of the current website is better than the proportion of conversion of the new website.

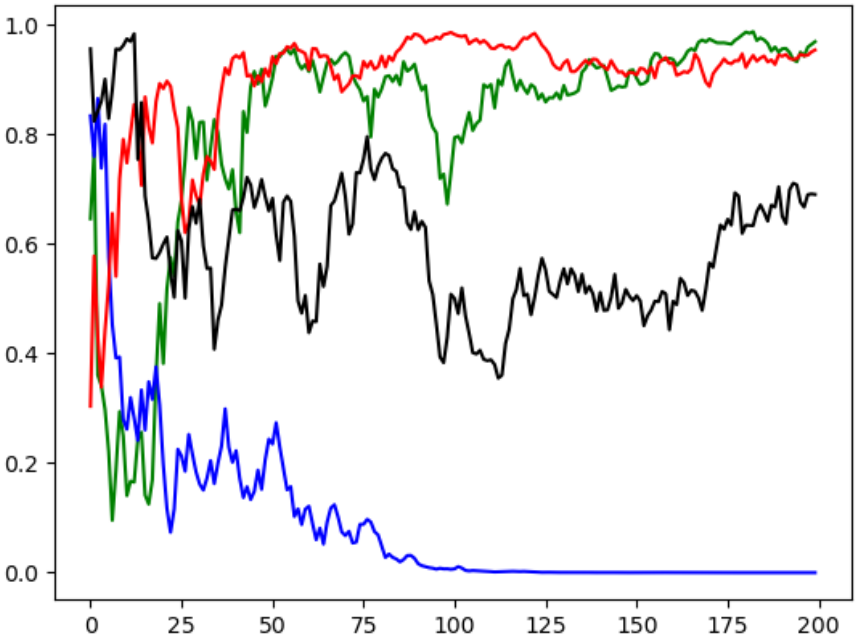


Figure 1. Monte Carlo simulation of the proportion of conversion of the current website is greater than the proportion of conversion of the new website of different scenarios compared to the actual data. The green line represents the simulation for the actual data collected. The black line represents an artificial simulation where both proportions of conversion are equal, or both 0.10. The red line represents an artificial simulation where the proportion of conversion of the new website is 0.10 and the proportion of conversion of the old website is 0.11. Lastly, the blue line represents an artificial simulation where the proportion of conversion of the new website is 0.11 and the proportion of conversion of the old website is 0.11.

In conclusion, our analysis suggests that the new website won't be as effective at encouraging customers to sign-up for the grocery club membership compared to the current website. We suggest follow-up studies with customers to determine why some customers sign up and why others don't. In particular, it could be useful to examine the motivations of the customers who were directed to the new website but chose to instead use the old one; finding where the sales pitch succeeded in these customers could be key to figuring out why the pitch didn't land with the other customers.

To: Insurance Companies in America
Subject: What Kind of People You Should and Shouldn't Insure

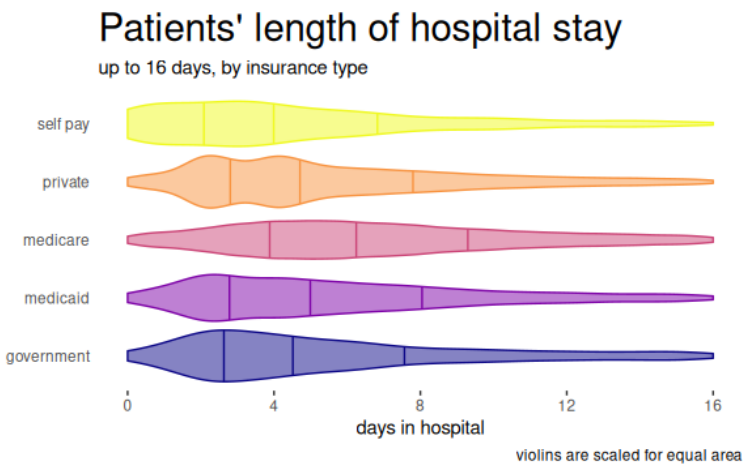
As an insurance company, you would like to be able to predict how long your clients would stay at the hospital in case of medical issues. You might want to avoid insuring people who are more likely to spend long periods at the hospital, as the cost of these long stays is a drag on company revenue. Therefore, our team of researchers and data analysts have come together to help you answer the question: what kind of people should you insure?

We used the MIMIC3d (Multiparameter Intelligent Monitoring in Intensive Care) data provided by MIT [1]. This dataset contains 58977 samples of patients and 28 attributes. The following is a list of the attributes, what they mean, and their values.

Attribute	**Explanation**	**Values**
hadm_id	Patient identification.	100001, 100003, etc.
gender	Gender of patient.	M (male) or F (female).
age	Age of patient, in years.	0, 14-89.
LOSdays	How many days the patient stayed at the hospital.	0.0, 6.17, 14.38, etc.
admit_type	The type of admission to the hospital.	'ELECTIVE', 'EMERGENCY', 'NEWBORN', or 'URGENT'.
admit_location	The location from which the patient was admitted to hospital.	"EMERGENCY ROOM ADMIT", "HMO REFERRAL/SICK", etc.
AdmitDiagnosis	Diagnosis with which the patient was admitted to hospital.	"CHEST PAIN", "PNEUMONIA", "OVERDOSE", "NEWBORN", etc.
insurance	Patient's medical insurance type.	"Government", "Medicaid", "Medicare", "Private", or "Self-Pay".
religion	Patient's religion.	"CATHOLIC", "MUSLIM", "NOT SPECIFIED", "OTHER", etc.
marital_status	Patient's marital status.	"SINGLE", "MARRIED", "DIVORCED", etc.
ethnicity	Patient's ethnicity.	"WHITE", "PORTUGUESE", "ASIAN - JAPANESE", etc.
Num*, TotalNumInteract	Billing for callouts, diagnosis, procedures, CPT events, inputs, labs, micro labs, notes, outputs, Rx, procedure events, transfers, chart events, and the total hospital bill payable by the patient.	0, 33.75, 49.69, etc.

AdmitProcedure	Procedure that will be performed on the patient.	“Vaccination NEC”, “Hemodialysis”, “Opn aortic valvuloplasty”, etc.
LOSgroupNum	A group number assigned based on the patient’s length of stay.	0, 1, 2, or 3.
ExpiredHospital	Whether the patient passed away during their hospital stay.	Takes the value “0” (did not pass during hospital stay) or “1” (passed during hospital stay).

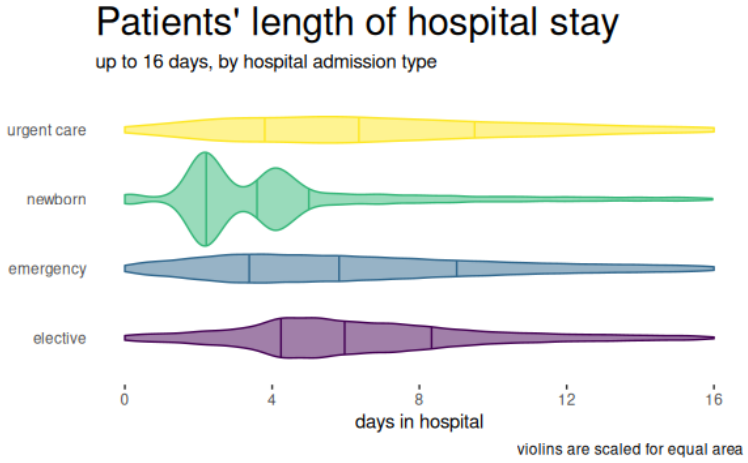
Note: We could not find proper documentation of the attribute definitions, so some of the explanations are educated guesses (in particular, *Num**, *TotalNumInteract*, and *ExpiredHospital*).



We investigated the length of patients’ hospital stays, as this is often a large factor that comes into play in hospital bills, and the length of stay directly determines how long any individual patient will have to take time off from work. Notably, the type of insurance the patient possessed (or didn’t possess) seems to have a substantial association with the length of a hospital stay. There are without a doubt confounding factors – Medicaid and Medicare recipients, for example, tend to be older or in poorer health, and this tends to lengthen any stay in hospital – but there are also likely some direct consequences of insurance type.

For example, in the self pay (no insurance) category, the number of days in hospital is concentrated toward the low end of the scale. This is at least partially because people without insurance are unlikely to have the means to afford longer hospital stays. Curiously, people with private insurance seem to conclude their hospital stays around the 3- or 5-day marks. Is this a consequence of the fine print in insurer agreements, or are patients with private insurance more prone to certain types of procedures with quick recovery times?

The type of a patient’s hospital admission also had some influence on the number of days spent in hospital. Unsurprisingly, this is chiefly the case among newborn admissions. As a general rule, a newborn infant admitted to hospital care will be discharged around the 3rd or 4th day of their stay, likely due to the need for monitoring after a minor treatment. This is a much earlier expected discharge than for other hospital admission types.



To visualize the distribution of LOSdays, we created a histogram and counted the amount of values in 100 bins. So for example, the first bar is a count of how many patients stayed at the hospital between 0.0 and 2.95 days.

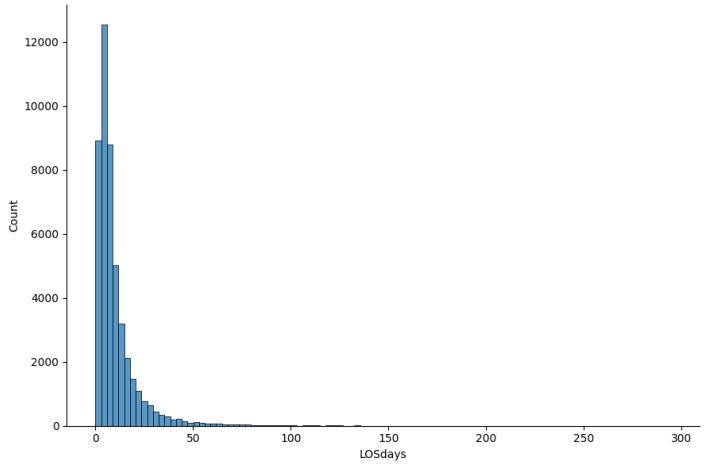


Figure 1. LOSdays Distribution

The vast majority of patients stay at the hospital for a short period of time. The probability of a patient staying an extra day decreases exponentially after the second bin (after about 6 days). To find a formal probability distribution, we fit 5 popular distributions to the data and took the one with the lowest sum of squared errors as the best fit. The sum of squared errors is a measure of the total deviation of the data from the fitted line, where the differences between the data points and the fitted line are squared and then summed up. The best fit was the Burr distribution, with a sum of squared error of 0.000039. The second-best was the log-normal distribution with a sum of squared error of 0.000233.

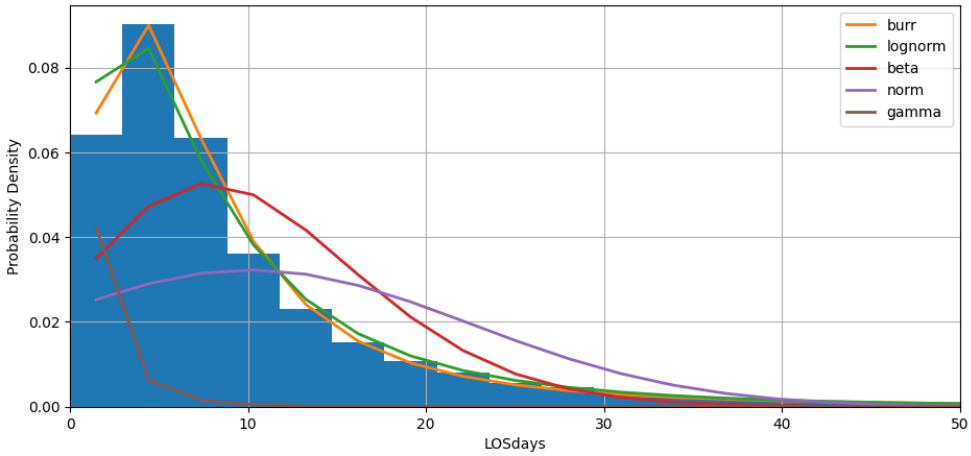


Figure 2. Five statistical distributions fitted to the LOSdays data.

We also tested an additional 75 distributions, and Burr was still the distribution with the lowest sum of square error, strongly suggesting that this is the best distribution to fit the data. This distribution does in fact make sense for the LOSdays attribute. The Burr distribution is often used to model data with heavy tails [2]. For example, it is good at modeling income data, where most people have an income within a certain range, but a small proportion of the population earns significantly more [3]. This is analogous to the LOSdays data. Most patients stay at the hospital only for a few days, but there are also

patients that stay at the hospital for a much longer time. For example, there are 3 patients in the dataset that have stayed at the hospital for more than 200 days.

Next, we sought out to see precisely how close the LOSdays distribution is to the normal distribution. We created a Q-Q plot (quantile-quantile plot) which plots the quantile of the normal distribution against the quantile of the LOSdays distribution (see Figure 3a). It compares the density of the normal distribution at a certain position on the x-axis to the density of the LOSdays distribution at the same position.

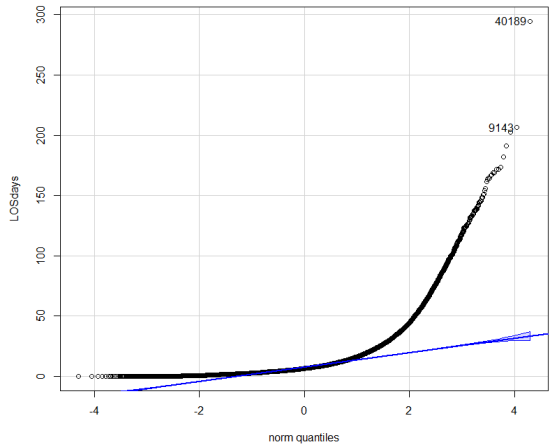


Figure 3a. Q-Q plot of normal distribution against the LOSdays distribution.

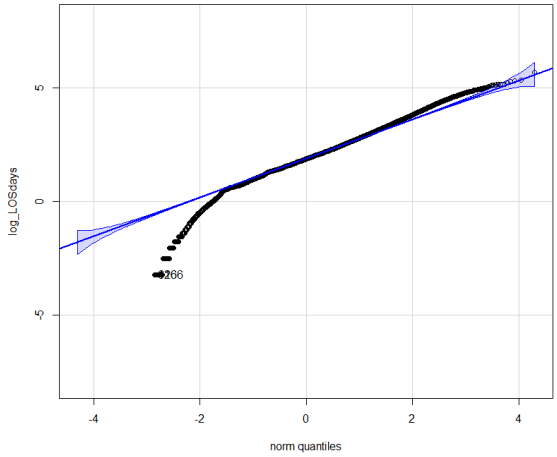


Figure 3b. Q-Q plot of normal distribution against the LOSdays distribution.

If the LOSdays distribution were normal, the data points would have been within or close to the light blue area around the dark blue line in the graph on the left. Evidently this is not the case — the data points seem to instead have an exponential shape. More specifically, in the right tail of the normal distribution, the density of the LOSdays distribution is higher than the normal distribution (the points are above the blue line). This is because there are more values in the right tail of the LOSdays distribution than there would be for a normal distribution. For example, there are 15 values in LOSdays that are above 150. Therefore, to normalize the data, we decided to apply a natural logarithm transformation (inverse of exponential function). That is, we took the natural logarithm of the data points and then plotted its quantile against the normal distribution quantile again (see Figure 3b).

We see that the logarithm of the LOSdays distribution is more similar to the normal distribution, although still not exactly normal. This also agrees with the results of the distribution fits from earlier. The second best distribution was the log-normal distribution, which is a probability distribution of a random variable whose logarithm is normally distributed [4]. Furthermore, given that the logarithm of the length of stay follows a normal distribution, we can use that to calculate the probability that a patient has a length of stay greater or equal to 2 days, which is 0.885.

In order to know which kinds of patients have the shortest lengths of stay in hospitals, a Bayesian Linear Regression was run to study the impact that different characteristics have on length of stay. The variables for which the coefficients were significantly not zero were the following:

Variable	**Impact on Length Of Stay**
age	Increase of 0.006340013 days for every year of age.
Admit type is emergency	Increase of 0.5372575 days if admit type is emergency.

Variable	Impact on Length Of Stay
Admit type is newborn	Increase of 0.2092496 days if admit type is newborn.
Admit type is urgent	Increase of 0.5683122 days if admit type is urgent.
Admit location is physician referral or normal delivery	Increase of 0.5999942 days if admit location is physician referral or normal delivery.
Admit location is transfer from another hospital	Increase of 0.584074 days if admit location is transfer from another hospital
Admit location is transfer from another health facility	Increase of 1.691234 days if admit location is transfer from another health facility
Patient died during their stay	Decrease of 0.1894158 days if patient dies during their stay
Admit type is elective	Default value for the Admit type variable, causes no changes to length of stay in this model.

The fact that patients who passed during their stay at the hospital spent less time than average in hospital is not particularly surprising. However, the fact that patients who transferred from a different health facility spent almost two days longer (on average) in hospital is striking. Could these patients have been transferred due to the presence of specialist facilities/doctors at the receiving hospital? Are only the most severely impacted patients receiving these transfers? More analysis is needed to be certain. However, we should keep in mind that in the case of your clients being transferred between health facilities, you can expect a longer and more expensive stay.

In conclusion, to minimize the length of stay of patients they insure, insurance companies should focus on providing coverage for elective procedures and admissions where the customer is the most likely to die. Furthermore, they should avoid other types of admissions such as emergency, newborn and urgent, avoid insuring older customers, and avoid insuring customers when they get transferred between facilities.

References

Used in the report:

1. "The Medical Information Mart for Intensive Care." *MIMIC*, mimic.mit.edu/. Accessed 26 Oct. 2023.
2. A R Hakim et al. "Properties of Burr Distribution and Its Application to Heavy-Tailed Survival Time Data." *IOPscience*, 2021, iopscience.iop.org/article/10.1088/1742-6596/1725/1/012016.
3. "Burr Type XII Distribution." *MathWorks*, www.mathworks.com/help/stats/burr-type-xii-distribution.html. Accessed 26 Oct. 2023.
4. "Log-Normal Distribution." *Wikipedia*, Wikimedia Foundation, 2 Oct. 2023, en.wikipedia.org/wiki/Log-normal_distribution#:~:text=The%20log%2Dnormal%20distribution%20is,ln(X)%20are%20specified.

Used in our analysis:

Raoniar, Rahul. "Finding the Best Distribution That Fits Your Data Using Python's Fitter Library." *Medium*, The Researchers' Guide, 9 Sept. 2022, medium.com/the-researchers-guide/finding-the-best-distribution-that-fits-your-data-using-pythons-fitter-library-319a5a0972e9.

Bastos, P. M. de. (2022, June 22). *How to use bayesian inference for predictions in Python*. Medium. https://towardsdatascience.com/how-to-use-bayesian-inference-for-predictions-in-python-4de5d0bc84f3