BZAN 535 - Homework 6

Kevin Gardner

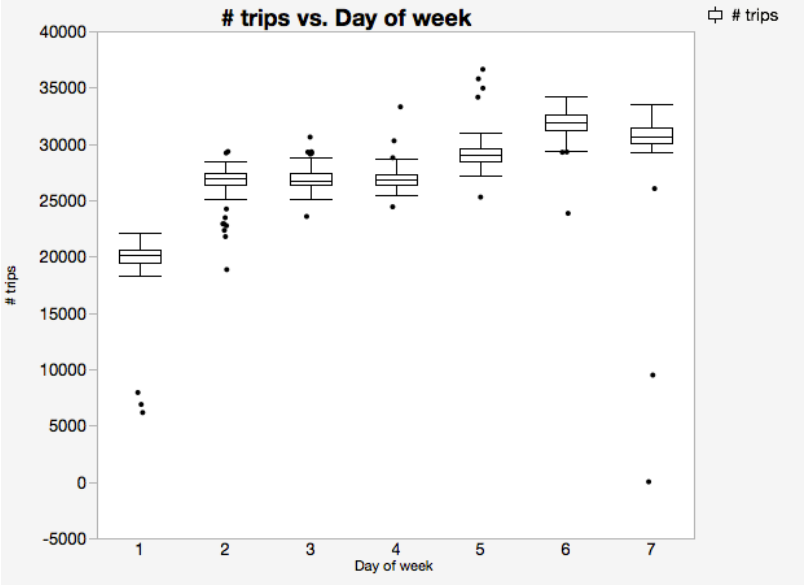
10/05/2016

Question 1

The file Kroger_trips_grouped_by_date_hw6_2016.jmp contains spending at Kroger over a 15-month period (April 2010 – June 2011) for 100,000 households,GROUP BY date.

a. Present a graph of side-by-side box plots for the number of households shopping by day of the week. Write a paragraph explaining the insights from this graph.

From the boxplots we can see that the number of trips generally increase throughout the week. The number of trips is greatest on Friday and Saturday. Sundays and Mondays appear to have the widest overall variation. The number of trips are very similar on Monday, Tuesday and Wednesday, then there is a moderate increase in the number of trips on Thursday.



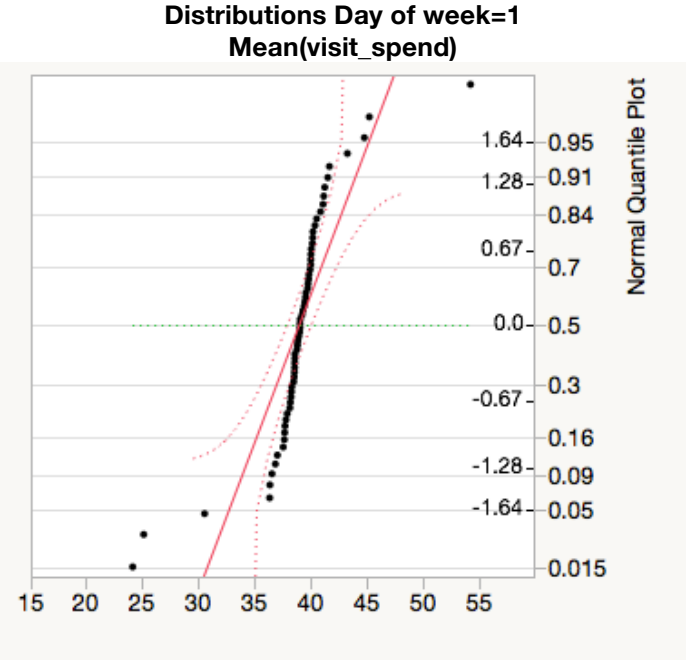
b. In what way do the results differ from what you would expect, based on your experience of grocery shopping? Do you think Kroger has modified the data – or might this be accurate for some subset of Kroger shoppers? (This question necessarily requires some speculation. Day 1 corresponds to Sunday.)

Based on my experience of grocery shopping I expected the largest number of trips on Friday and Saturday and a medium number of trips in the middle of the week. I also expected the largest number of trips to be on Sunday. I suspect that either my experience or this subset is not representative of most households. The latter is probably the case since the subset would only be comprised of Kroger card members.

Question 2

This question (and #3) concerns the mean dollars spent per visit for each of the 445 days.

a. Averages tend to be normally distributed. For Day of week = 1, produce a normal quantile plot for mean spend. This will reveal 7 outliers. Report the p-value for the Goodness of Fit test with and without these 7 values. State the null and alternative hypotheses, and your conclusion about this distribution.



Fitted Normal (With outliers)
Parameter Estimates

Type	Parameter	Estimate	Lower 95%	Upper 95%
Location	μ	39.012954	38.067999	39.957909
Dispersion	σ	3.7829588	3.2223151	4.5816184

-2log(Likelihood) = 350.928957830046

Goodness-of-Fit Test

Shapiro-Wilk W Test

W	Prob<W
0.725790	<.0001*

Note: Ho = The data is from the Normal distribution. Small p-values reject Ho.

Fitted Normal (Outliers Removed)
Parameter Estimates

Type	Parameter	Estimate	Lower 95%	Upper 95%
Location	μ	39.109144	38.770308	39.44798
Dispersion	σ	1.2770071	1.0781192	1.5665642

-2log(Likelihood) = 188.634174834273
Goodness-of-Fit Test
Shapiro-Wilk W Test

W	Prob<W
0.985310	0.7162

Note: Ho = The data is from the Normal distribution. Small p-values reject Ho.

With outliers included we get a p-value of <.0001*, so we reject the null hypothesis and conclude the data is not from the normal distribution. With outliers removed we get a p-value of 0.7162, so we fail to reject he null hypothesis and conclude the data is from a normal distribution.

c. Examine all the outlier points and comment on what the unusually large mean spend dates have in common, and on what the unusually small spend dates have in common.

The outlier points generally correspond with days surrounding Christmas (New Year's Eve) and Easter. Average spend appears to be very small on these specific days and very large in the preceding days.

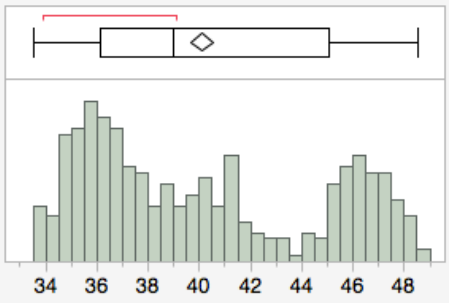
Question 3

After excluding outliers in 2a and 2b, you will have about 400 days left.

a. Present a histogram for mean spend from all of the days of week together. Describe the resulting distribution. What is the standard deviation for this sample of about 400 values and why is it larger than the standard deviations for individual days?

Distributions

Mean(visit_spend)



Quantiles

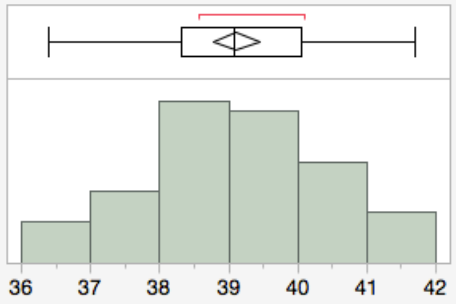
100.0%	maximum	48.521858172
99.5%		48.502169806
97.5%		48.071282008
90.0%		46.855948244
75.0%	quartile	45.084764079
50.0%	median	38.966148833
25.0%	quartile	36.160685351
10.0%		34.98989611
2.5%		34.000257612
0.5%		33.530525274
0.0%	minimum	33.516545758

Summary Statistics

Mean	40.106229
Std Dev	4.5327212
Std Err Mean	0.2266361
Upper 95% Mean	40.551779
Lower 95% Mean	39.660679
N	400

Distributions Day of week=1

Mean(visit_spend)



Quantiles

100.0%	maximum	41.716004645
99.5%		41.716004645
97.5%		41.651731375
90.0%		40.968207849
75.0%	quartile	40.05214807
50.0%	median	39.085171951
25.0%	quartile	38.30629939
10.0%		37.485414191
2.5%		36.390533913
0.5%		36.379246324
0.0%	minimum	36.379246324

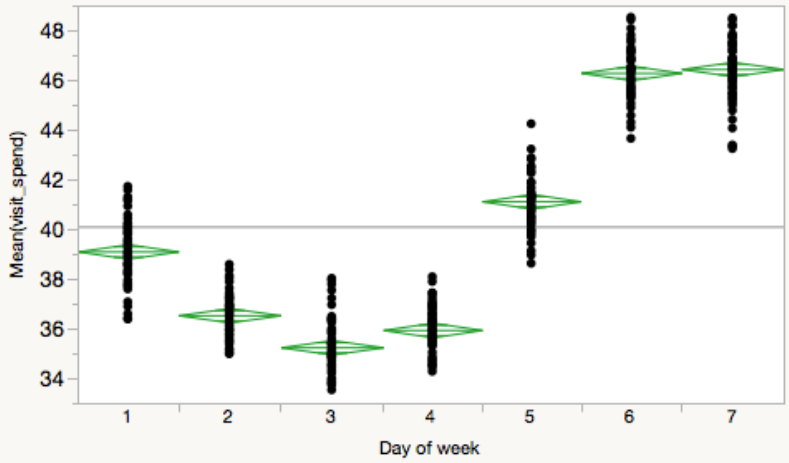
Summary Statistics

Mean	39.109144
Std Dev	1.2770071
Std Err Mean	0.1691437
Upper 95% Mean	39.44798
Lower 95% Mean	38.770308
N	57

The distribution of mean spend for combined days is multimodal and slightly right skewed. The standard deviation for combined days and for Day of week=1 are 4.5327212 and 1.2770071, respectively. The former is larger because it includes the variation within each day of the week as well as variation across different days of the week. As we determined from part 1a, there appears to be variation in mean spend between days of the week, so including this additional source of variation results in a larger standard deviation.

b. Construct an ANOVA using X = Day of Week, Y = mean spend. Report the F statistic and p-value. What hypotheses is this testing? Summarize the conclusion for this test in the language of the problem.

Oneway Analysis of Mean(visit_spend) By Day of week



Excluded Rows 45

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Day of week	6	7702.6158	1283.77	1019.105	<.0001*
Error	393	495.0631	1.26		
C. Total	399	8197.6790			

The ANOVA tests the hypothesis

H0: mean spend is equal across all days of week

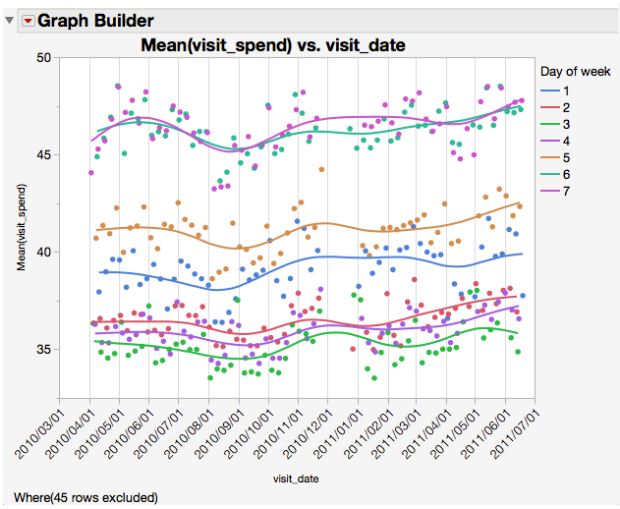
H1: mean spend is not equal across all days of week

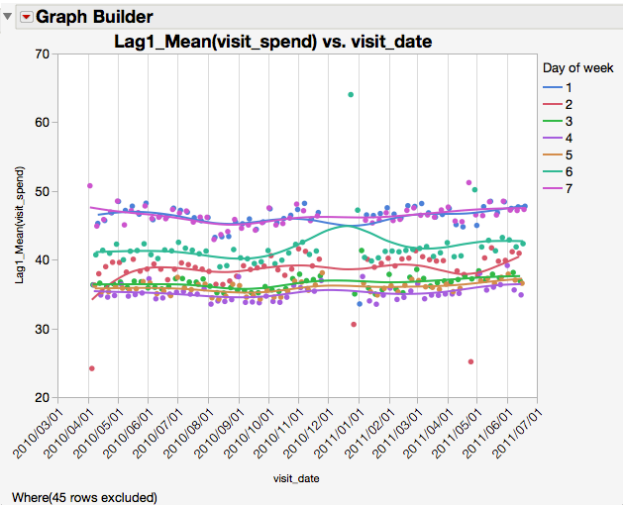
From the result we see the observed F Ratio = 1019.105 >>1. The p-value is <.0001*, so the differences between days is highly significant and we reject H0 in favor of H1.

c. This same ANOVA can be obtained by using Fit Model and putting Day of Week as the only term in the model, and mean spend as Y. As we did in HW5, check for autocorrelation using the Durbin-Watson test. Report the autocorrelation value. To what do you attribute this positive autocorrelation? To help you answer this question, create a plot using Graph Builder taking X = Date and overlay using Day of Week. Interpret this graph and the autocorrelation.

d. Durbin-Watson

Durbin-Watson	Number of Obs.	AutoCorrelation	Prob<DW
0.6740287	400	0.6554	<.0001*





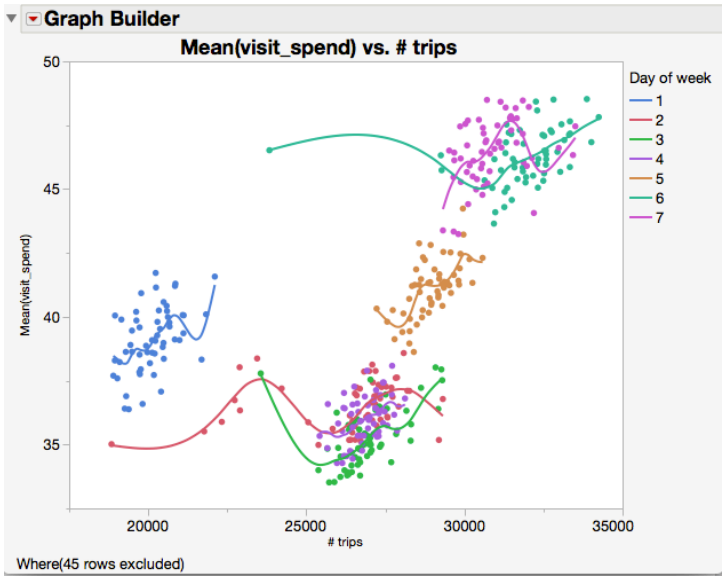
According to the Durbin-Watson test there is an autocorrelation of 0.6554. The corresponding p-value is <.0001*, so we reject the null hypothesis and conclude there is positive serial correlation.

In the graph we that mean spend, represented by the vertical axis, depends on day of the week. But we also see that mean spend on for a particular day might depend on mean spend the day before. If this is true, then the wavelike behavior in mean spend over time should flatten out if plot the 1-day lag instead of contemporaneous mean spend. The result is shown in the plot on the right. It is not obvious (to me) why there would be positive serial correlation. But it means that the visit spend on a given day of week depends on visit spend the previous day.

Question 4

Are days that have more shoppers than usual for that day of the week prone to have higher dollars per basket (than is typical for that day of the week)?

a. Produce a graph using Graph Builder with mean dollar spend on the yaxis, number of trips on the x-axis, and use Overlay for Day of Week. Display this graph and use it to answer the question above.

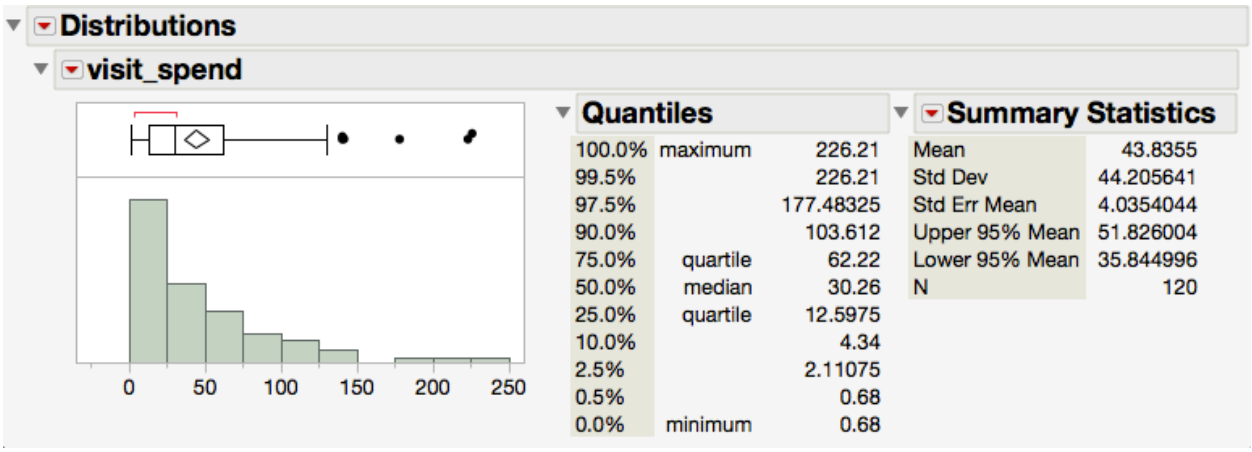


From the graph we observe that days of the week with highest # trips also seem have higher mean spend. This is illustrated by day 3 and day 7, which appear at in the far upper right of the graph.

Question 5

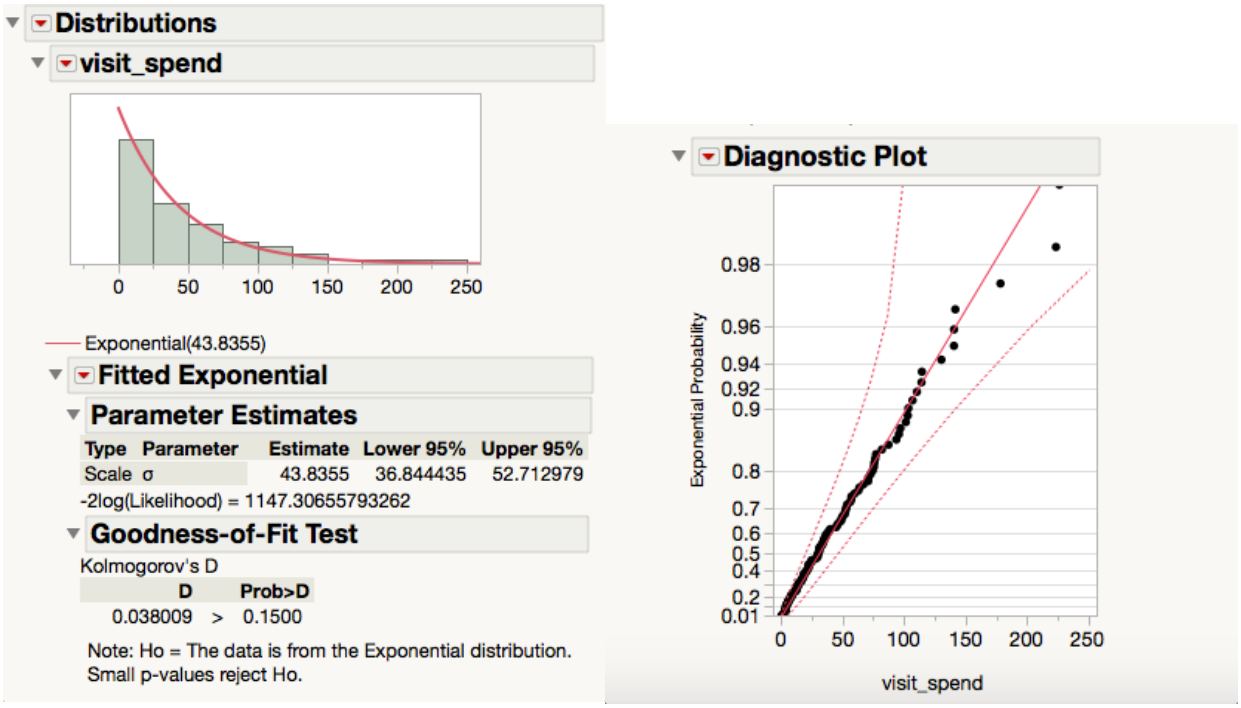
Given that there are slightly more than 12 million trips, the individual trip data are a bit overwhelming. So I post just a sample of 120,000 trips (a 1% sample).

a. Using JMP's Table->Subset, select a random sample of just 120 rows from this file of daily trips. Using Distribution, obtain a histogram for visit_spend with this sample. Report the mean and standard deviation for this distribution. Save this sample for question 5b.



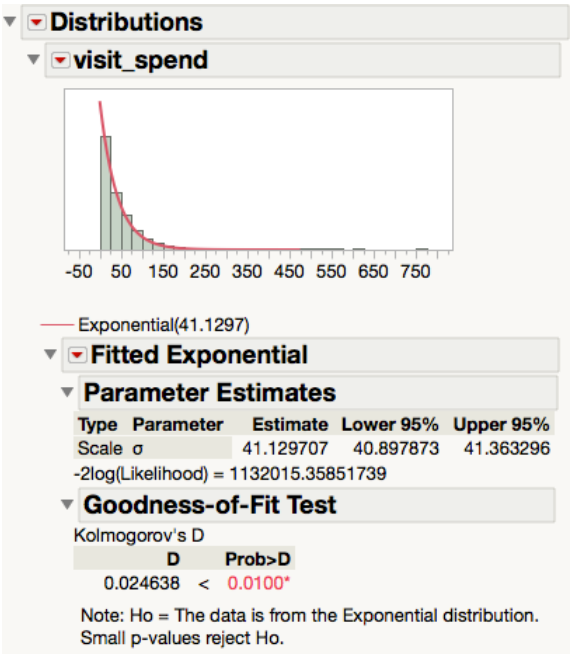
For the random sample of 120 trips, we observe a mean of 43.8355 and standard deviation of 44.205641.

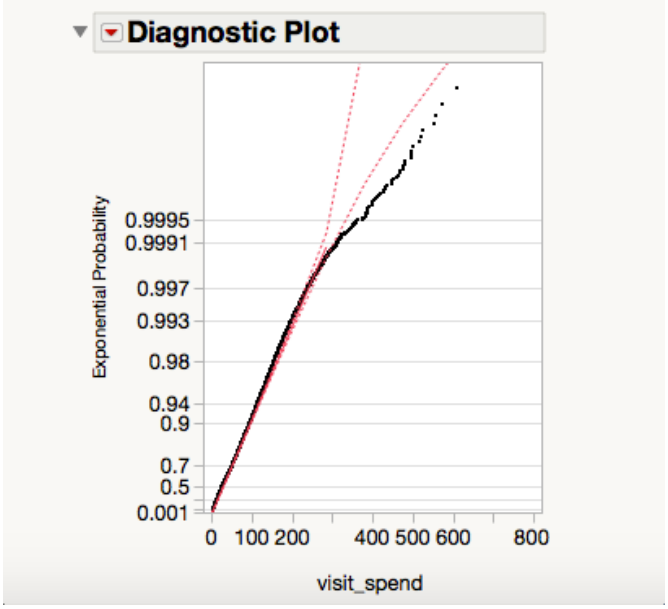
b. The mean and standard deviation are equal for the exponential distribution. Use "Continuous Fit", under Analyze -> Distribution in JMP, to see whether the exponential distribution is a suitable model for the amount spent per day by this population of families. Show and interpret the goodness of fit test and the diagnostic (quantile) plot.



The Goodness-of-fit test gives a D-value of 0.1500, so we fail to reject the null hypothesis and conclude that the sample is from the Exponential distribution.

c. Now return to the original data file of 120,000 rows. Show the histogram, diagnostic quantile plot, and goodness of fit test for exponential distribution fit for the same variable in 5b.



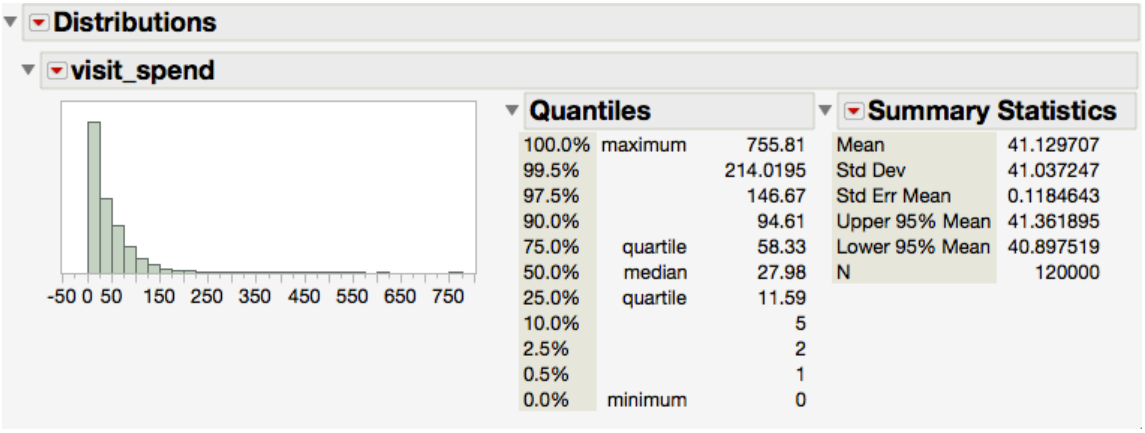


d. Explain what you learn from the difference in the goodness of fit tests in 5b (where n=120) and 5c (where n = 120,000).

I learned that the Durbin-Watson test is not very robust to sample sizes that are small relative to the population of interest. In this case we only had a sample size of 1%, so it would be interesting to see how large the sample size must be in order to reject the null hypothesis with reasonable consistency.

Question 6

Using the 120,000 rows, report an 80% prediction interval for sales at Kroger. Explain why your chosen method (t-based or quantiles-based) was the most appropriate.



Since the sample size is large we can construct our prediction interval using the quantiles given above. Thus we obtain the 80% prediction interval [5, 94.61]. This is means we are 80% certain that trip spend for a randomly selected observation will between $5 and $95.61.

Question 7

I took a simple random sample of 120,000 rows – a 1% sample. Suppose instead that I had sampled all trips for 1000 randomly selected households (1% of households). The resulting sample would be about the same size. Why is it better to sample 1% of households and have all their trips? What more could we have learned about shopping behavior that way?

Having all of the trips for 1000 randomly selected households would allow us to pair the observations and make inferences based on data that are less noisy. For example, if we wanted to analyze the effect of a promotion then we could look at purchase behavior before and after to perform more direct comparisons. There are a ton of variables that might affect the overall sales for an item that happens to be on promotion. But having the capability to pair data at the household-level would hopefully allow us to hold a lot of those variables constant and detect the signal that we're interested in.