Case 107: Fish Story - Not Too Many Fish in the Sea

Dr. DeWayne Derryberry, Idaho State University
Department of Mathematics

Fish Story: Not Too Many Fish in the Sea

Paired t-Test and Nonparametric Tests

Key ideas: Adjusting for Inflation, Additive Versus Multiplicative Models, Hypothesis Testing, Matched Pairs, Paired t-Test, Nonparametric Tests.

Background

Seafood exists in the open seas, which are owned by no one. Because of this some economists believe that seafood is harvested in numbers much higher than is optimal. Although the theory is quite complicated, the basic idea is simple. Since no one owns or manages the open seas, those who fish cannot gain, in the long run, by refusing to harvest today. If I refuse to harvest fish today, the stock of seafood will not grow, because someone else will harvest the fish if I don't. (These ideas, discussed in detail in microeconomics, are a discussion of public goods versus private goods). Because of this, over time, seafood is overharvested and becomes increasingly scarce.

If supplies of seafood are dwindling, seafood will becomes more expensive, relative to other goods, over time.

The Task

Use the DASL Fish Prices data to investigate whether there is evidence that overfishing occurred from 1970 to 1980.

Source: DASL (The Data and Story Library), lib.stat.cmu.edu/DASL/Datafiles/FishPrices.html

The Data Fish Story.jmp

The data table includes seafood prices (cents per pound) in 1970 and 1980 for 14 types of seafood.

Type	Type of Seafood
1970 Price	Price of fish in 1970
1980 Price	Price of fish in 1980
Adjusted 1970 Price	The 1970 price adjusted for inflation using the CPI (1970 Price x 2.1237)

Seafood has risen in price in the period from 1970 to 1980, but just about everything increased in price between 1970 and 1980 (just ask your parents). In fact, economists have a way of quantifying this general increase (called inflation), using the consumer price index. The consumer price index (CPI) is based on the cost of a representative bundle of consumer goods in various years. Using a typical online CPI calculator it is possible to determine that $100 worth of consumer goods in 1970 would cost $212.37 in 1980. Therefore, if seafood prices increased at the same rate as other consumer goods in the period from 1970 to 1980, prices would typically increase by a factor of 2.1237.

Analysis

If overfishing has occurred in the time period, the price of fish will rise faster than the overall price of all consumer goods. On the other hand, if fish are about as plentiful in 1980 as they were in 1970, the typical price will be about the same in each period after the CPI adjustment.

Our hypotheses are:

> Ho: The mean price of seafood is about the same in 1970 and 1980 after adjusting for inflation.
> Ha: Mean seafood prices were higher in 1980, than in 1970, even after adjusting for inflation.

or

> Ho: $\mu_{diff} = 0$ versus Ha: $\mu_{diff} > 0$ where μ_{diff} is the mean of the differences.

Multiplicative Models and Logarithmic Transformations

Before we perform the appropriate test and assess whether there is support for our claim, we need to consider whether we should work with the data on the original scale or take a logarithmic transformation of the data.

Why would we consider the logarithmic transformation?

Of course, an examination of the data may suggest a logarithmic transformation (skewness, the presence of outliers, etc.). In this case, there are a priori reasons for believing such a transformation would make sense. The basic question is whether we think an additive model (no transformation) or a multiplicative model (logarithmic transformation) is best.

Since prices generally change as %age rather than by some absolute amount, a multiplicative model makes sense. For example, if gas prices go up by 10%, it would not be surprising if milk prices were to go up by 10%. CPI is already based on this idea. In contrast, an additive model would suggest that if a gallon of gas went up by $0.25, we would expect a gallon of milk to go up by $0.25.

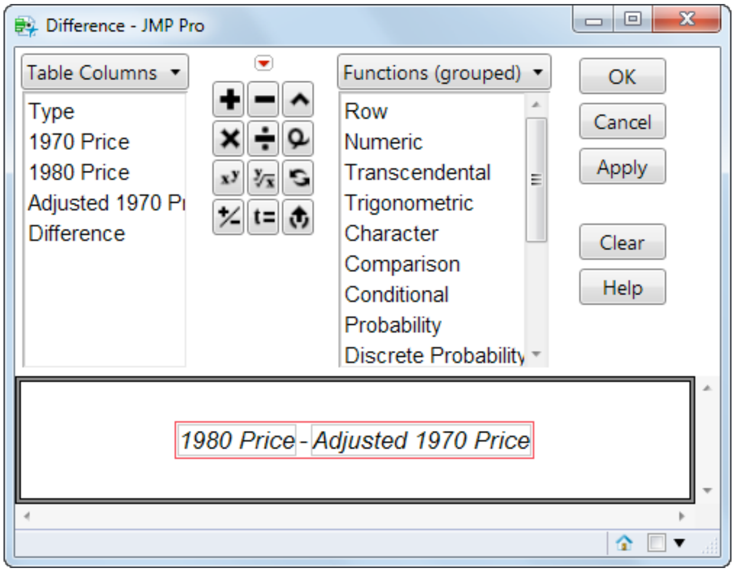
Given these facts, when we examine the data, we will have a predisposition to work with data on the logarithmic scale, unless the data screams for a different analysis.

Matched Pairs

Paired analysis occurs when two measurements are taken that are expected to be highly positively correlated, rather than independent. Soft shelled clam prices, in 1970 and in 1980, for example, are obviously strongly positively correlated. Examples of matched pairs include: measurements taken on identical twins, measurements of the same quantity over time, or two measurements taken on the same subject or other experimental unit.

With matched-pairs data, a difference is taken, producing a single measurement for each pair (presumably, the pairs are independent of each other). Then, a test is performed on (or confidence interval constructed from) the differences. For this reason, it is the difference that must satisfy the conditions for any test to be appropriate (rather than the original paired data).

Exhibit 1 Creating a Column of Differences



1980 Price - Adjusted 1970 Price

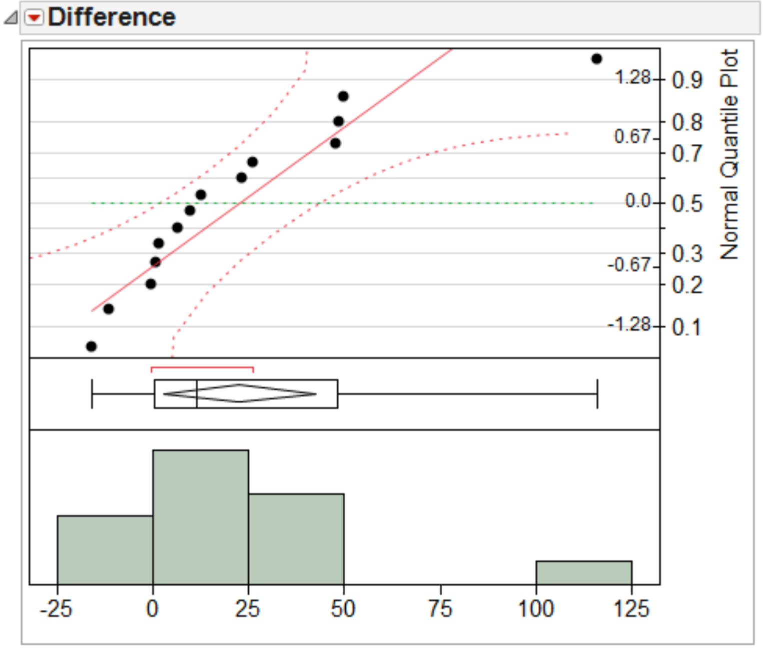
*(Create a new column in the data table, and rename it **Difference**. Right click on the column header, and select Formula to open the Formula Editor. To create the formula:*

1. *Select **1980 Price** from the columns list*
2. *Select the minus sign on the key pad*
3. *Select **Adj 1970 Price** from the columns list*
4. *Click OK.*

Note that this formula can also be created directly from the data table. Select the two columns in the data table, right-click and select New Formula Column > Combine > Difference.)

Although the data set is small, it appears that the distribution of differences is somewhat skewed and/or has a potential outlier (Exhibit 2). This confirms the value of a logarithmic transformation (multiplicative model).

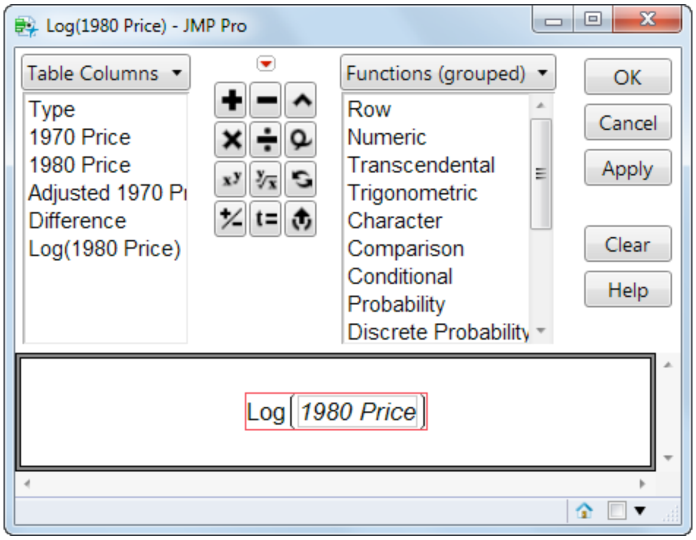
Exhibit 2 Distribution of Differences



*(Analyze > Distribution; select **Difference** as Y, Columns and click OK. From the red triangle, select Normal Quantile Plot.)*

Since we're interested in paired differences, we first transform the pricing data (Exhibit 3). Then, we create a column of differences for the transformed data.

Exhibit 3 Taking a Log Transformation



*(Create a new column in the data table, and rename it **Log(1980 Price)**. Right click on the column header, and select Formula to open the Formula Editor. To create the formula:*

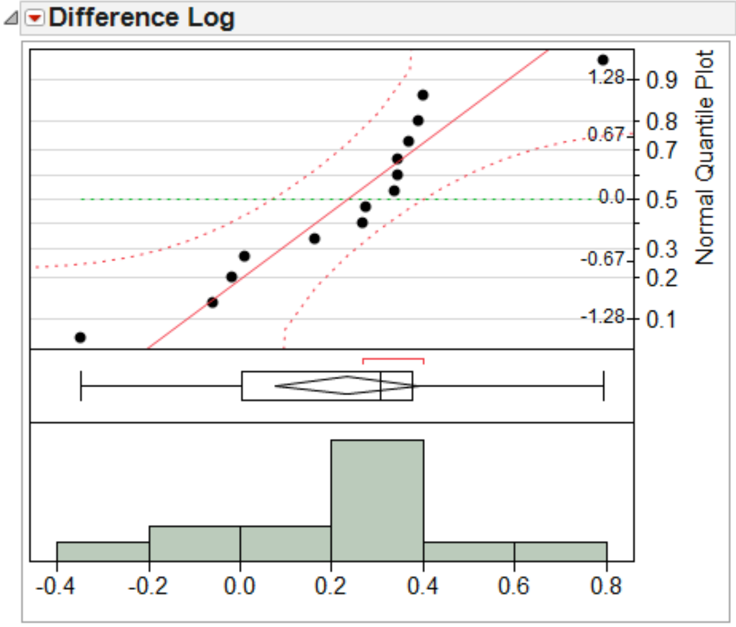
1. *Select **1980 Price** from the columns list*
2. *From the Functions (grouped) list select Transcendental, Log*
3. *Click OK.*

To create this formula column directly from the data table, right click on the column and select New Formula Column > Transform > Log.

*Repeat to create **Log(Adj 1970 Price).**)*

The differences of the logarithms of prices (which is also the logarithm of the ratio of prices), look quite good (Exhibit 4). There is little evidence of either skew or outliers.

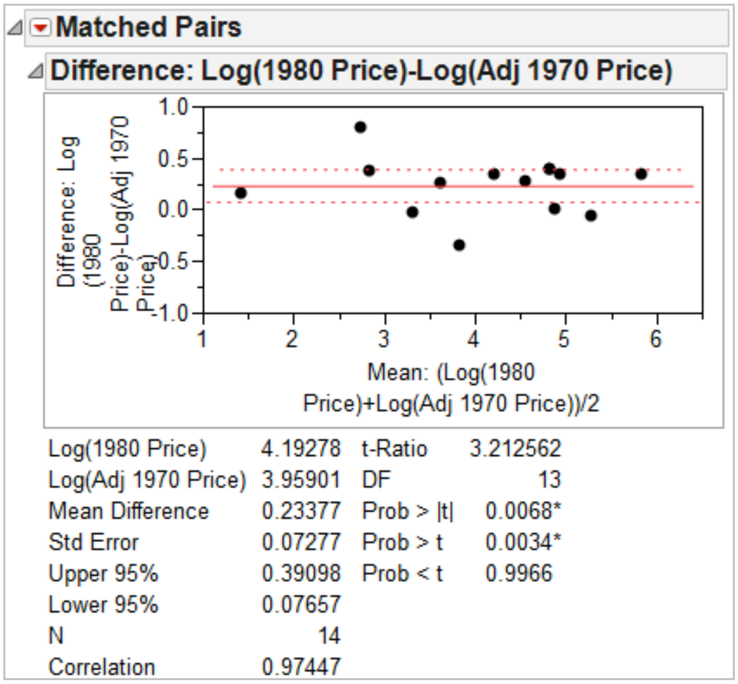
Exhibit 4 Differences for the Transformed Data



Paired t-Test

A paired t-test will be used to see if there is evidence of an increase in prices, above and beyond inflation, from seafood prices in 1970 to 1980. Note that a paired t-test is equivalent to performing a one sample t-test on the column of differences.

Exhibit 5 A Paired t-Test for the Transformed Data


Matched Pairs
Difference: Log(1980 Price)-Log(Adj 1970 Price)
Difference: Log (1980 Price)-Log(Adj 1970 Price)
1.0
0.5
0.0
-0.5
-1.0
1
2
3
4
5
6
Mean: (Log(1980 Price)+Log(Adj 1970 Price))/2

```
Log(1980 Price)      4.19278  t-Ratio    3.212562
Log(Adj 1970 Price)  3.95901  DF               13
Mean Difference      0.23377  Prob > |t|   0.0068*
Std Error            0.07277  Prob > t     0.0034*
Upper 95%            0.39098  Prob < t     0.9966
Lower 95%            0.07657
N                         14
Correlation          0.97447
```

(Use Analyze > Matched Pairs; select the two log price variables from Select columns, and click Y, Paired Responses. Click OK.)

The one sided *p*-value of 0.0034 (next to Prob > t) provides strong evidence that seafood prices have risen faster than inflation in the period under study (Exhibit 5). The correlation between the transformed 1980 and adjusted 1970 prices is 0.97447, confirming the appropriateness of the matched-pairs approach

The confidence interval provides an estimate of how much seafood prices have gone up (beyond inflation. Since it's in a log scale, it requires inverse transformation.

- 95% CI for Difference in Log Prices: (0.07657,0.39089)
- 95% CI for Difference in Prices: (1.08, 1.48)

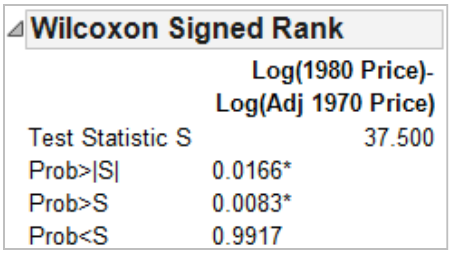
(Hint: To apply an inverse transformation in JMP, right-click on the numerical output in Matched Pairs and select Make into Data Table. Then, create a new column, and use the formula editor to transform Column 2 using the "Exp" function.)

So, even after accounting for inflation, seafood prices appear to have increased by 8% to 48%.

Wilcoxon Signed Rank Nonparametric Test

For those particularly cautious about violations of the assumptions, a nonparametric alternative is available. The Wilcoxon signed rank test assumes the data is symmetric about the hypothesized median of zero, which is milder than the assumption that the data is normal with a mean of zero.

Exhibit 6 Results for a Wilcoxon Signed Rank test



```
Wilcoxon Signed Rank
                              Log(1980 Price)-
                           Log(Adj 1970 Price)
Test Statistic S                        37.500
Prob>|S|           0.0166*
Prob>S             0.0083*
Prob<S             0.9917
```

(From the Matched Pairs output window, select Wilcoxon Signed Rank from the red triangle.)

The nonparametric test (Exhibit 6) leads to a similar conclusion. With a *p*-value of 0.0083, there is strong evidence of a shift upward in *median* seafood price (after adjusting for inflation) from 1970 to 1980.

Both the paired t-test and the Wilcoxon signed rank test produce very small *p*-values and strong evidence of an increase in seafood prices (beyond inflation) from 1970 to 1980. Which results should be reported? This decision should be based on how well the assumptions are met, rather than judging by results.

Summary

Statistical Insights

For completeness we showed both the parametric paired t-test and the nonparametric Wilcoxon Signed Rank procedures, but a professional analysis would involve a decision about which procedure to use early on in the process.

General considerations when choosing a procedure:

- Some procedures are incorrect in certain situations.

- For the most part, some procedures are just better than others in a given situations. A procedure is better when it makes more efficient use of the data – more narrow confidence intervals, and tests more likely to reject the null hypothesis when it is false while retaining the correct rejection rate when the null hypothesis is true.

Procedures that make more assumptions use the data more efficiently when the assumptions are met, but can produce poor results when the assumptions are not met. However, the notion that the assumptions are met is not as black and white as it sounds. How normal do data need to be in order to call them normal? And how much can you tell about normality from a small data set?

One statistician would argue that they began with sound reasons for taking a logarithmic transformation of the data, then data supported this transformation, and the final data looked quite normal, so a t-test was used. Another statistician would argue that, no matter what a priori reasoning is involved, the data set is too small to properly assess the assumption of normality, so a nonparametric test is best.

This is not really mathematics, it is modeling philosophy, and as such there are no correct answers. The best you can do is gain experience, develop intuition, and formulate (and periodically revise) an approach you can justify and live with.

Implications

What can we say, based on this analysis? We certainly can say these data suggest seafood prices are increasing relative to other consumer goods, but we cannot say this is due to overfishing. In order to say that overfishing is the issue we need three ingredients: a theory as to why overfishing would lead to the observed phenomenon (which I have sketched out a bit), many data sets (not just one), or other empirical evidence that seafood prices have indeed increased at a rate faster than inflation. And, we need to rule out other possible explanations. This is an observational study and there are lots of reasons seafood prices may be rising faster than other goods. For example, maybe advertising or the changes in the demographic composition of the population has increased the demand for seafood.

Is the sample representative? The seafood chosen is not a random sample of all possible seafood. We must consider whether it is a representative sample, even if it is not a random sample. One way to think about this is whether we would have gotten different results if we had chosen different seafood? While I think the sample is representative, I will leave the final judgment to experts on seafood economics.

JMP® Features and Hints

In this case study we used the Distribution Platform and normal quantile plots to visually assess the shape of distributions. We used the formula editor to create log-transformed data and create columns of differences. The Matched Pairs platform was used to conduct paired t-tests and nonparametric Wilcoxon signed rank tests.

Exercises

This exercise involves data on the labor force participation rate of women from 1968 to 1972 for 19 major US cities.

Source: From the DASL (The Data and Story Library), the US Bureau of Labor Statistics, http://lib.stat.cmu.edu/DASL/Stories/WomenintheLaborForce.html

1. Import the data into JMP from the website above, and save the file as Labor Rates.jmp.

2. Explain why these are paired data.

3. Compute the differences and assess normality.

4. Take the logarithmic transformation of these data and take the differences for the logged data. Assess normality.

5. Perform a paired t-test for both the original and the transformed data. Interpret the results, and describe the change with confidence intervals.

6. Perform an appropriate nonparametric test.

How do the *p*-values from these and the paired t-tests compare? Which result are you most confident in, and why?

7. To verify that a paired t-test is identical to a one sample t-test on the column of differences, use the Distribution platform and perform a one sample t-test and a signed rank test on the differences for the logged data.

   Compare the *p*-value and confidence intervals for the difference with those obtained in part 4.

SAS Institute Inc. World Headquarters +1 919 677 8000

JMP is a software solution from SAS. To learn more about SAS, visit **www.sas.com**
For JMP sales in the US and Canada, call 877 594 6567 or go to **www.jmp.com**