**Tips for the AP Exam/Formula Sheet**

**Create a flow chart/tree diagram detailing how to choose the appropriate inference procedure. Here are the 9 tests we have learned:**

1. **One sample z-test for a proportion**
2. **One sample t-test for a mean**
3. **Two sample z-test for a difference in proportions**
4. **Two sample t-test for a difference in means**
5. **Matched pairs t-test for a mean difference**
6. **Chi-square test for goodness of fit**
7. **Chi-square test for homogeneity of proportions**
8. **Chi-square test for independence**
9. **Linear Regression t-test for slope**

**Note: All of the procedures above (except the chi-square tests) have confidence intervals as well.**

**Modeling Non-Linear Data**

When statisticians look at a scatterplot, they often use the phrases “signal” and “noise.” The signal is the underlying form of the data and the noise is the random variation from that form.

For example, when we studied the relationship between fat and calories in hamburgers, we saw that hamburgers with more fat has more calories (that’s the signal) but even hamburgers with the same amount of fat can have different amounts of calories (that’s the noise).

If the residual plot looks randomly scattered, then our model has captured the whole signal and only the noise remains. If the residual plot shows a pattern, however, then our model has missed some of the signal and we should seek a better model.

**Therefore, when deciding which model to use, the most important thing to do is consider the residual plot for each model.**

Here are some data showing the highest individual baseball salary (in millions of dollars) for various years between 1980 and 2001.

|  |  |
| --- | --- |
| Years Since 1980 | Salary ($ millions) |
| 0 | 1 |
| 2 | 2.04 |
| 9 | 3 |
| 10 | 4.7 |
| 11 | 5.3 |
| 16 | 8.5 |
| 17 | 11 |
| 18 | 12.5 |
| 19 | 15 |
| 21 | 25.2 |

Note: We often shift the data on the x-axis so it is closer to the origin. This makes the computations and interpretations easier in many cases. In this case, x = 17 means year = 1997.

Sketch a scatterplot of this data with the LSRL. Make a residual plot and discuss if the linear model is appropriate.

* *

**

*Since there is an obvious pattern in the residual plot, it is clear that our model has not captured the signal of the data. That is, our model does not have the right form.*

In general, there are 2 approaches for finding models for non-linear data.

1. Fit a curve to the data.
2. Make the data straight and fit a line.

For several reasons, the second method is preferred in AP Statistics, although in the real world, option 1 is often preferred.

To straighten the data, we can employ all sorts of transformations: squaring, square rooting, finding the reciprocal, etc. In AP Stat land, our transformation of choice will usually be the logarithm. We can use either the base 10 logarithm (log) or the natural logarithm (ln) and get similar models. In general, we will use ln since this is what most statisticians prefer, but you should be able to use both.

Since the original data seems to look like exponential growth, lets take the natural log of the salaries and make a scatterplot of year vs. ln(salary).



Since the transformed data looks fairly linear, find the LSRL for these data and make a residual plot.



Note: Our new model is in the form: . Models in this form are called \_\_Exponential Models\_\_\_\_\_\_\_\_\_\_\_\_\_ since they can be re-expressed as .

The residual plot still shows a possible U shaped curve, but it is much more scattered than the linear model. Thus, the exponential model will be more useful than the linear model. Remember, there is no perfect model, but some are more useful than others.

Use this model to predict the highest salary for the years 1993 and 2005. Do these seem reasonable?

Note: This method only works when the y-values are positive (you cannot find a log of 0 or a negative number) and start close to 0. If you need to shift the y-values, remember to shift back when making predictions!

Here are some data about our solar system:

|  |  |  |
| --- | --- | --- |
| Planet | Distance from Sun  (million miles) | Length of Year  (Earth years) |
| Mercury | 36 | 0.24 |
| Venus | 67 | 0.61 |
| Earth | 93 | 1 |
| Mars | 142 | 1.88 |
| Jupiter | 484 | 11.86 |
| Saturn | 887 | 29.46 |
| Uranus | 1784 | 84.07 |
| Neptune | 2796 | 164.82 |
| Pluto | 3666 | 247.68 |

Sketch a scatterplot with the LSRL and make a residual plot. Does the linear model seem appropriate?

 



Transform the data by taking ln (y). Sketch a scatterplot with the LSRL and make a residual plot. Does the exponential model seem appropriate?

 



Transform the data by taking ln (x) and ln (y). Sketch a scatterplot with the LSRL and make a residual plot. Does this model seem appropriate?

 



Note: Models in the form  are called \_Power Models\_\_\_\_\_\_\_\_ since they can be re-expressed as .

In 2002, researchers discovered another object orbiting the sun 4 *billion* miles away from the sun. About how long should it take to orbit the sun?

**Conclusion:**

There are many more transformations we can try when logarithms don’t work. Sometimes taking a square root will help straighten out a data set. Sometimes squaring a data set or using the reciprocals of a data set will work.

Besides transforming data to make it linear, you can also use many types of mathematical functions in an attempt to fit the data. Some of these can be found in the stat:calc menu on the TI-83 and many others can be found on computer software.

The final decision about which model to use should be based on the residual plots. The model whose residual plot has the most random scatter does the best job of capturing the true form (signal) of the relationship.