Bonus 4 - 14 - 10 **Due in class 4-19-10**

Sometimes we see an apparent pattern in data, but perhaps not a straight line. In a surprising number of cases simply transforming x and/or y using logarithm or other functions will reshape the plot into a nearly straight line. If so, least squares offers a convenient and simple way to produce such a line. In your chapter 10 readings for today the focus is on doing this, without much regard for the statistical aspects of so doing. The statistics is taken up (in chapter 27) next week.

All references in chapter 10 to

(1) having residuals plot in a formless way, or
(2) having the normal probability plot of residuals look like a line,
are really asking if things appear to be consistent with 2D normal behavior.

That is great if it happens but actually many important uses of "change scales of x and/or y then use a straight line fit" seem to work very well as descriptions of data without either of (1) or (2). The world just produces lots of interesting examples of this.

I've chosen an interesting example for you to look at: Zipf's Law.

A. Look up two references to Zipf's Law on the web. Find and cite two that you can understand and make sure that they agree.

B. The picture below is from **Statistics, A guide to the unknown., Tanur et. al., Wadsworth, 1988**. It almost tells the story of Zipf's Law. Consider the 20 most populous metropolitan areas as of 1980.

Take x = Log[metro rank], y = Log[metro population]). New York City ranked 1 with population ~20 million and Los Angeles ranked 2 with population ~15 million. Zipf's Law suggests a roughly straight line for the 20 points (x, y):

	x = Log[rank]	y = Log[population]	
New York	Log[1] = 0	Log[20] = 2.99573	
Los Angeles	Log[2] = 0.69	Log[19] = 2.94444	etc.

The picture accomplishes this by using log scales on each of the horizontal and vertical axes, so if you have the log-log paper you actually enter x = rank and y = population size in millions to view the log-log plot which should appear roughly like a straight line.

To determine the slope and point of means for that line you do have to process the scores (Log[x], Log[y]).

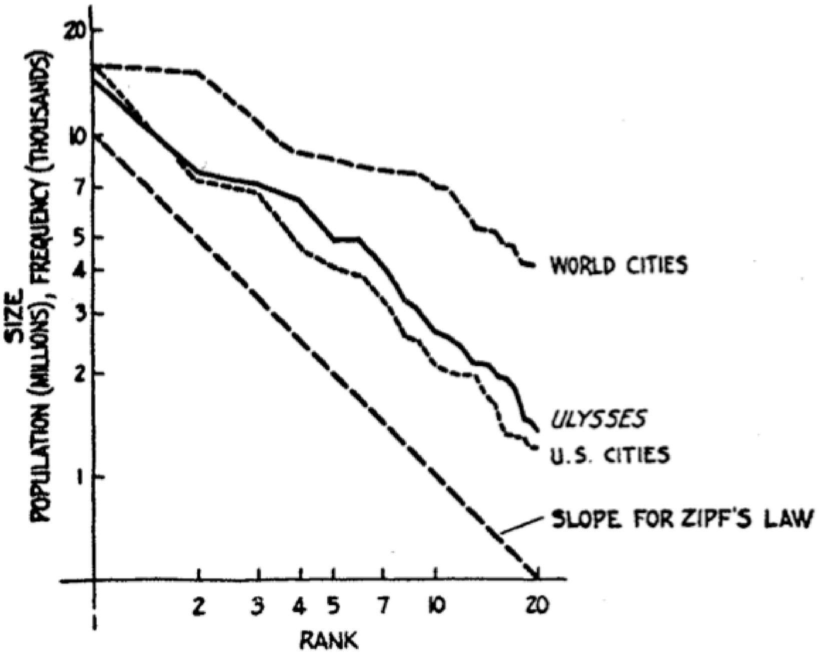
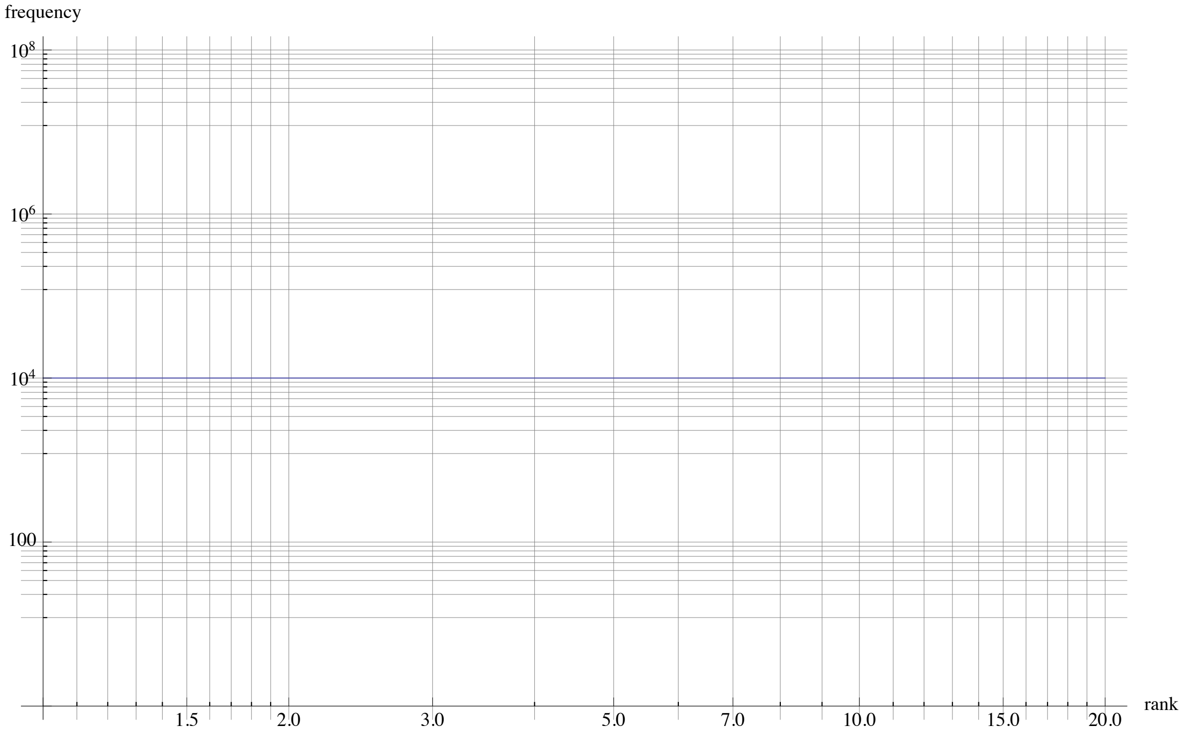


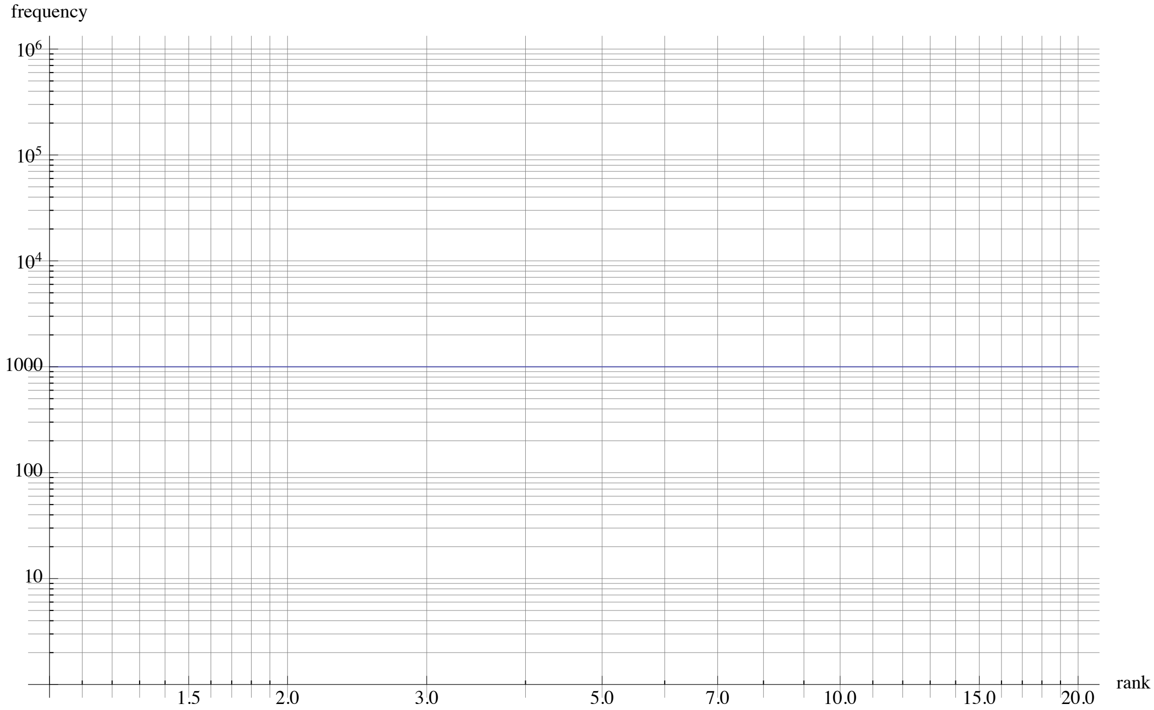
Figure 1 *Logarithm of size plotted against logarithm of rank for frequencies*

Your assignment is to gather data for which you think that Zipf's Law may apply. Use at least 10 points. Here are some images of log-log plots that you will find useful as you work on this bonus.

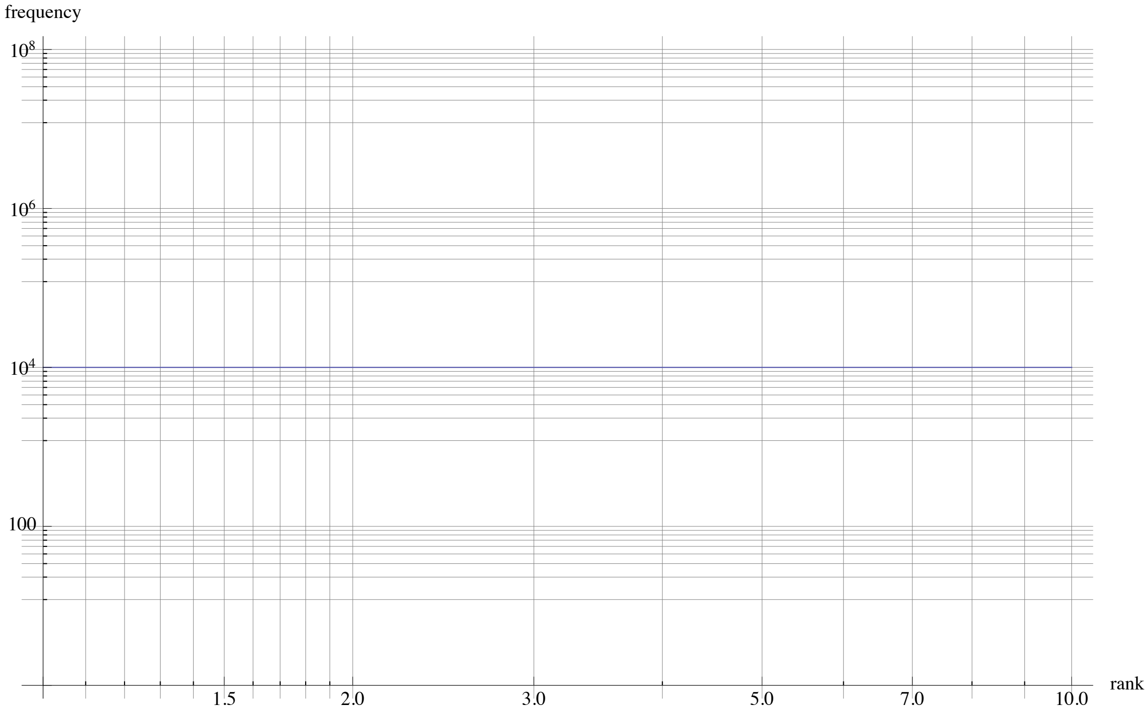
```
LogLogPlot[10 000, {x, 1, 20}, AxesLabel → {rank, frequency}, GridLines → True]
```



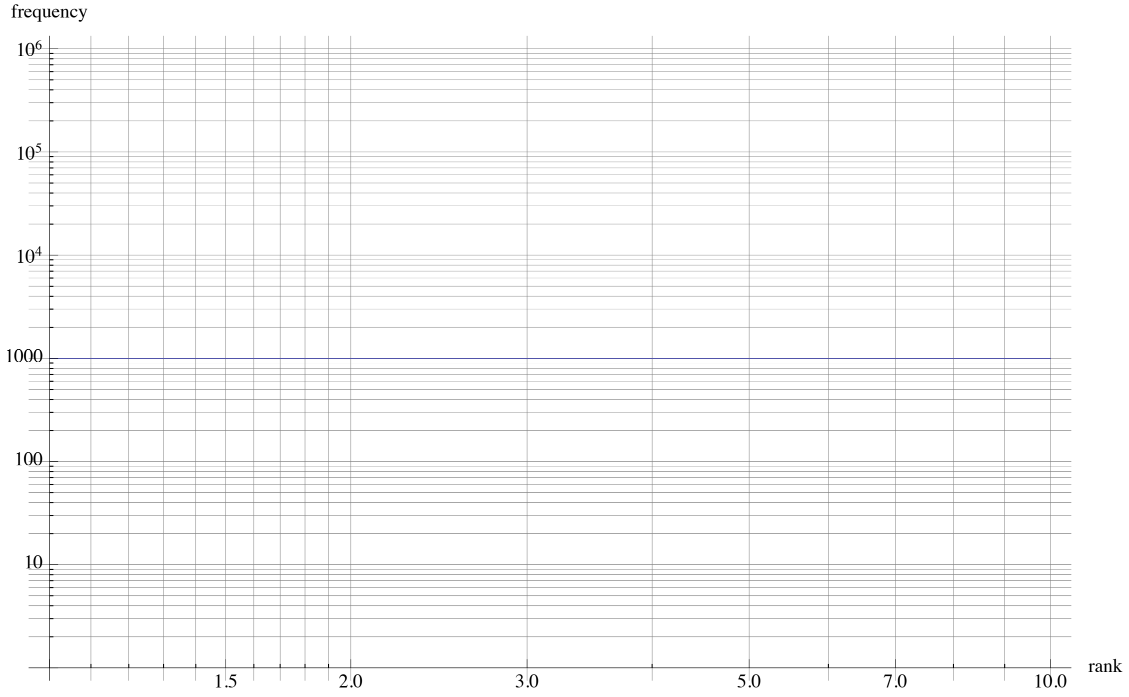
```
LogLogPlot[1000, {x, 1, 20}, AxesLabel → {rank, frequency}, GridLines → True]
```



```
LogLogPlot[10 000, {x, 1, 10}, AxesLabel → {rank, frequency}, GridLines → True]
```



```
LogLogPlot[1000, {x, 1, 10}, AxesLabel → {rank, frequency}, GridLines → True]
```



- **(Added late) Compare the standard scale plot of (Log[rank], Log[frequency]) for the Ulysses data shown below) with what you see when you plot the regular data (rank, frequency) in the log-log paper. In your opinion, is log-log paper plot faithfully representing the degree to which the plot of (log[rank], log[frequency]) is ~ straight line in regular scale?**

James Joyces' Ulysses word counts are given below.

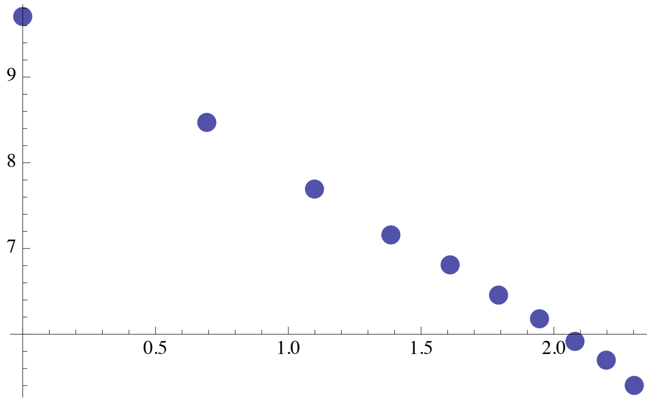
```
In[1]:=  joyce = {16 432, 4776, 2194, 1285, 906, 637, 483, 371, 298, 222}

Out[1]=  {16 432, 4776, 2194, 1285, 906, 637, 483, 371, 298, 222}
```

Here is a regular scale plot of (Log[rank], Log[frequency]) for the Ulysses

```
In[2]:=  ListPlot[Table[{Log[i], Log[joyce][[i]]}, {i, 1, 10}], PlotStyle → PointSize[.03]]
```

Out[2]=


9
8
7
0.5
1.0
1.5
2.0