

APM 153 LECTURE FIFTEEN – Transforming Data, Basic Stats, Intro to Excel

Log Transformations

(1) Often, data needs to be transformed before trying to calculate any statistical measures. Why? Because before transformation, the data are not easily compared.

(2) Take for example the raw scores from our first exam. The raw scores ranged from 133 to 270 points. **Based on the points alone, how do we assign a grade (A, B, C, etc)?**

(3) We transform the data by dividing by the total number of points possible which gives us a **percentage**. In the case of our first exam, the total number of points was 270 which means that the raw percent scores ranged from 49.3% to 100%.

(4) It is much easier to first transform the data then compare values. Same thing with statistics. And in some cases, data must be transformed before they can be compared.

(5) One way of transforming the data is to use the **log** or **log10** commands.

(6) Transforming the data using log10 or the natural log forces the data to conform to a standard scale.

(7) Sometimes, we only want to transform either the x or the y values but not both. If we use the log10 or log functions to transform only the y or the x values we say that the resulting plot is a **semilog** plot.

Basic Statistics

(8) The most basic statistics we can calculate about a set of data are the so-called **measures of central tendency** – the **mean**, the **median**, and the **mode**.

(9) The **mean** of course is **the average** of all the values. Given,..

2 4 6 8 10 the average = 6

(10) The **median** is **the middle value** if all the values were sorted from high to low.

Again, given,.. 2 4 **6** 8 10 the median = 6

(11) The mode is more difficult to calculate, but easy to understand. The **mode** is the **most common value**. Given the data below, which value is the mode?

2 4 3 2 4 3 4 4 2 4 3 2 4 4 4 5 4 the mode = _____

(12) The three “measures of central tendency” are just three ways of saying that values tend to cluster around the average.

(13) In addition, we also can calculate how closely the values cluster around the average by calculating the **sum of the squares**, the **variance**, and the **standard deviation**.

(14) The **sum of the squares** is the sum of the differences between each value and the mean squared. For example, given the values in 8 and 9 above,

VALUE	MEAN	DIFFERENCE	SQUARED
2	6	4	16
4	6	2	4
6	6	0	0
8	6	2	4
10	6	4	16

SUM OF THE SQUARES = 40

(15) Another way of thinking about the sum of the squares is that it is the total of how much individual values differ from the mean.

(16) The variance therefore can be thought of as the average amount that individual values differ from the mean. The variance can be calculated as,...

VARIANCE = SUM OF THE SQUARES / NUMBER OF VALUES

for our data VARIANCE = 40/5 = 8

(17) Finally, the **standard deviation** is the square root of the variance and **defines a range around the average** in which we **should find most of the values**.

Example Statistics from Our First Exam

mean	- 82.5%
median	- 83.3%
std dev	- 11.7
mode	- four people got a 90%
most common grade	- A Eleven people got an A on this exam.
range	- 49.3 – 100%

Hey, What About the Data for Assignment Six?

(18) If you still can't download the data from the APM 153 website, send me an email message and I will send you copies of SITE1.DAT, SITE2.DAT, and SITE3.DAT.

Intro To Excel

(19) Today's Lecture on Excel will be posted as a MS Powerpoint Presentation on the APM 153 course website.

