Biost 517: Applied Biostatistics I
Emerson, Fall 2009

Homework #2 Key
October 30, 2009

Written problems: To be handed in at the beginning of class on Wednesday, October 14, 2009.

*On this (as all homeworks) unedited Stata output is **TOTALLY** unacceptable. Instead, prepare a table of statistics gleaned from the Stata output. The table should be appropriate for inclusion in a scientific report, with all statistics rounded to a reasonable number of significant digits. (I am interested in how statistics are used to answer the scientific question.)*

Questions for Biost 514 and Biost 517:

1. The class web pages contain descriptions of two datasets
    - PSA data (psa.doc)
    - MRI and cerebral atrophy data (mri.pdf)

    a. For each of the described scientific questions, briefly characterize the type of statistical question to be answered. That is, using the classification presented in class, characterize the problem as clustering of cases, clustering of variables, quantifying distributions within groups, comparing distributions across groups, or prediction, identifying any variable whose distribution is of interest and any groups that might be being compared.

Answer:

- ***PSA dataset*: The scientific question is to decide whether time in remission is associated with nadir PSA value. This can be addressed by a statistical analysis which compares the distribution of times in remission (as measured by OBSTIME and INREM) across groups defined by nadir PSA value. Such an analysis must take into account the censoring indicated in the OBSTIME and INREM variables. A statistical analyst who did not know how to do such an analysis could alternatively consider using OBSTIME and INREM to determine who was still in remission at 24 months (the earliest censoring time) and compare the distribution of NADIRPSA across groups defined by whether they in remission at 2 years. In either case, to address the question of whether any such association is just a reflection of an association between bone scan score and time in remission, we would also have to make comparisons across groups that were similar with respect to BSS. To make the binary decision, we would use a test, but, again, we are undoubtedly also interested in point and interval estimates.**
- ***MRI dataset*: The scientific question is to decide whether time to death is associated with MRI findings believed to be indicative of cerebral atrophy and whether any such association is indicative of disease processes unrelated to other systemic disease. This**

question will ultimately compare time to death (as measured by OBSTIME and DEATH) across groups defined by levels of MRI atrophy (as measured by ATROPHY, WHGRD, NUMINF, and VOLINF). Such comparisons will be ideally made in such a way as to make the groups comparable with respect to demographic variables (AGE, MALE, RACE, WEIGHT, HEIGHT), the behavioral variables (PACKYRS, YRSQUIT, ALCOH, PHYSACT), the variables measuring concurrent disease (CHF, CHD, STROKE, DIABETES, GENHLTH), and markers of potential subclinical disease (cardiovascular disease risk LDL, SBP, WEIGHT; liver disease risk ALB, PLT; kidney disease risk ALB, CRT; bone marrow disease PLT, lung disease FEV; nutritional status ALB (note that many of the markers serve dual roles)) . We would primarily be interested in estimating the magnitude of any differences in survival as well as the precision with which those estimates are made.

b. For each of the datasets, classify the available measurements with respect to the statistical role they might play in answering the scientific question. That is, using the classification presented in class, identify which variables might be outcome measurements, predictors of interest, subgroup identifiers for interactions, potential confounders, precision variables, surrogates for the response, or irrelevant.

Answer:

- ***PSA dataset*: The outcome variable is OBSTIME and INREM (in combination they indicate measurements of time in remission subject to censoring), and the predictor of interest is NADIRPSA. The question of NADIRPSA having an association "independent of an effect" due to BSS or PS (sometimes referred to as NADIRPSA being an "independent risk factor") is common wording meant to consider the possibility that an observed association between NADIRPSA and time in remission was just due to the fact that NADIRPSA was associated with bone scan score or performance status, and that those latter variables were the true predictors of remission time. That is, such a question relates to whether the NADIRPSA and time in remission association might be confounded by BSS and PS. Thus, BSS and PS (and GRADE, AGE, and PRETXPSA, while we are at it) are potential confounders. PTID is irrelevant to the scientific question (though might be of interest statistically to ensure that no repeated measurements were made on any individual). (In this observational data set, any variable which might predict the outcome is automatically considered a potential confounder by me. It is not until I look at the data that I might be able to decide whether it is truly a confounder or merely a precision variable.)**
- ***MRI dataset*: The outcome variable is OBSTIME and DEATH (in combination they indicate measurements of time to death subject to censoring), and the predictor(s) of interest are the MRI characteristics ATROPHY, WHGRD, NUMINF, and/or VOLINF. The question of whether the MRI findings are merely indicative of other systemic disease or suggestive of a primary central nervous system disease will be addressed by also considering potential confounders that include the demographic variables (AGE, MALE, RACE, WEIGHT, HEIGHT), the behavioral variables (PACKYRS, YRSQUIT, ALCOH, PHYSACT), the variables measuring concurrent disease (CHF, CHD, STROKE, DIABETES, GENHLTH), and markers of potential subclinical disease (cardiovascular disease risk LDL, SBP, AAI, WEIGHT; liver disease risk ALB, PLT; kidney disease risk ALB, CRT; bone marrow disease PLT, lung disease FEV; nutritional status ALB (note that many of the markers serve dual roles)). (In this observational data set, any variable which might predict the outcome is automatically considered a potential confounder by me. It**

is not until I look at the data that I might be able to decide whether it is truly a confounder or merely a precision variable.) The variable DSST might also be an indicator of CNS disease, but as it is not directly a measure of MRI changes, it is not the predictor of interest. It is a variable that we would likely not use except to perhaps document associations between the subclinical MRI measures and the clinical signs of CNS impairment. The variables PTID and MRIDATE are largely irrelevant to the primary question of interest, though they might be of interest to assess issues related to repeated measurement on the same subject and or quality assurance over time.

c. For each of the datasets, classify the available measurements with respect to the type of measurement: qualitative versus quantitative, unordered versus partially ordered versus ordered, discrete versus continuous, and interval versus ratio.

Answer:

- **PSA dataset: Age, PSA nadir, and pretreatment PSA are continuous, quantitative variables measured on a ratio scale. Performance status is (to my mind) continuous and quantitative, but recorded discretely. I would probably regard it to be merely an interval scale, as it maxes out at 100, and I am not at all sure that ratios make any sense at all. Some people might regard that it is merely an ordered qualitative variable, as the levels of performance status are rather subjective. Bone scan score and grade are clearly ordered categorical (qualitative) variables. Patient ID is unordered categorical. Observation time is a continuous, quantitative variable measured on a ratio scale, but in terms of its scientific relevance, it is most important to note that it represents censored observations. INREM is a binary variable (discrete, ordered), but because it is measuring remission status over varying periods of time, it is important to note that the scientific interpretation of this variable requires knowledge about OBSTIME.**
- **MRI dataset: AGE, WEIGHT, HEIGHT, PACKYRS, YRSQUIT, ALCOH, PHYSACT, LDL, ALB, CRT, PLT, SBP, AAI, FEV, DSST, ATROPHY, VOLINF are continuous, quantitative variables measured on a ratio scale. I note that like performance status in the PSA dataset, ATROPHY maxes out at 100. NUMINF is a discrete quantitative variable. GENHLTH, CHD, STROKE, DIABETES, and WHGRD are ordered categorical variables. MALE and CHF are binary variables. RACE is a nominal variable. PTID is a nominal variable coded as a number. MRIDATE is a nominal variable measuring a continuous variable but in a coded form. OBSTIME is a continuous, quantitative variable measured on a ratio scale, but in terms of its scientific relevance, it is most important to note that it represents censored observations. DEATH is a binary variable (discrete, ordered), but because it is measuring vital status over varying periods of time, it is important to note that the scientific interpretation of this variable requires knowledge about OBSTIME.**

2. This problem deals with a data set containing various measurements made on a sample of generally healthy elderly adults. The primary goal in assembling this particular data set was to investigate the role of MRI findings in patient survival. The data (mri.txt) and documentation (mri.pdf) can be found on the class web pages. The file mri.txt can be downloaded and read into Stata using the command (typed all on one line)

infile ptid mridate age male race weight height packyrs yrsquit alcoh physact chf chd stroke diabetes genhlth ldl alb crt plt sbp aai fev dsst atrophy whgrd numinf volinf obstime death
using mri.txt

The questions can be answered using the Stata commands (other commands would also work)

- **tabstat** …, **stat(n mean sd min p25 med p75 max iqr r) col(stat) format**
- **hist** …, **bin(20)**
- **means …**

Note that I added the statistics "iqr" for interquartile range and "r" for range. You will have to use "means" to get the geometric mean, though it could also be obtained by generating a new variable that is the log transformed lab values, taking the mean of that new variable, and then exponentiating the result (you would do this last step with "display" or a hand calculator).

You many want to create a new variable which dichotomizes survival at the requested levels. There are many ways to do this. One way is as follows:

- **g**enerate *surv5yr = 1*
- **replace** *surv5yr* = 0 if *obstime < 5*365.25*

If you use this method, you will also need to make sure that missing data is handled appropriately. In this case, we can set all cases with missing data for *obstime* to also have missing for *surv5yr* (a period by itself is used as the code for missing data):

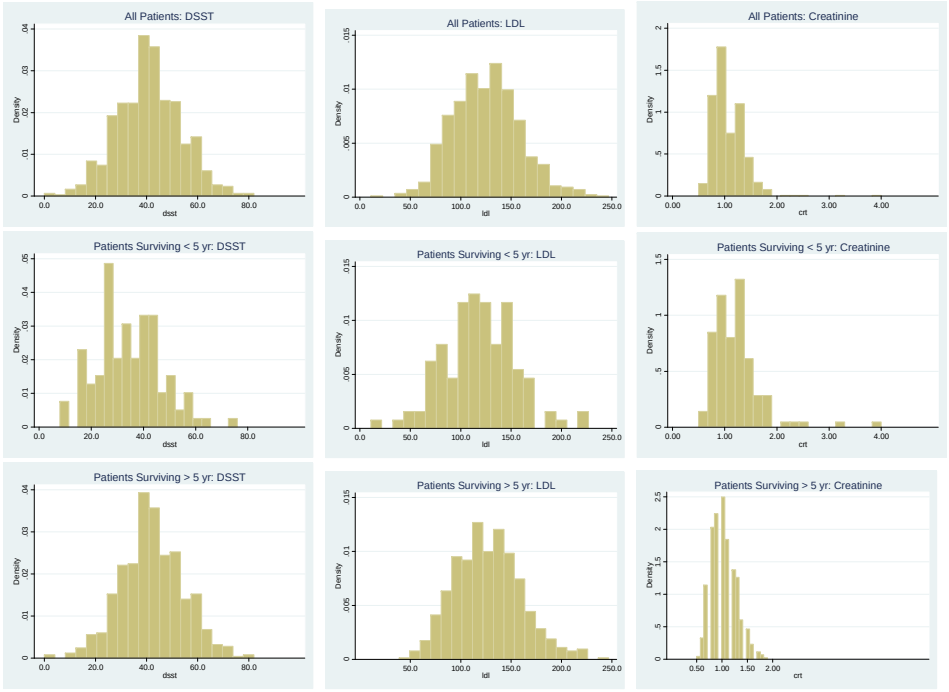
- **replace** *surv5yr* = . if *obstime* == .

Similar variables could be created to indicate low scores on the DSST (for this homework we define "low" as less than 30), high levels of LDL (which we will define as greater than 150 mg/dl), or high levels of creatinine (which we will define as greater than 1.4 mg/dl).

a. The variable *obstime* represents an incomplete measurement of the time from study enrollment to a patient's death. That is, for some patients, *obstime* contains the number of days between study enrollment and death, and for other patients *obstime* contains the number of days between study enrollment and "locking" of the database for data analysis. Such data is called "right censored", because when the variable *death=0,* we only know that the patient survived longer than the time recorded in *obstime*. We do not know the exact timing of the patient's death. In the prefatory remarks to this problem, I suggested that you create a variable *surv5yr* indicating whether a patient has survived at least 5 years. Why is this variable valid scientifically? Provide descriptive statistics justifying your answer.

b. Using the three laboratory values of LDL, creatinine, and DSST generate the following descriptive statistics for each group defined by whether or not they survived for 5 years:
   - Histogram
   - Number of cases with missing data

- Mean
- Geometric mean (only for LDL and creatinine—why?)
- Median
- Mode (it suffices to take an approximate mode from a histogram)
- Standard deviation
- Variance
- Minimum and maximum
- Range (the difference between minimum and maximum)
- 25^{th}, 75^{th} percentiles
- Interquartile range (the difference between 25^{th} and 75^{th} percentiles)
- Proportion of cases with "high" laboratory values (as defined above)



Above are the histograms for DSST (left column), LDL (middle column), and creatinine (right column) displayed for the entire sample (top row), those patients surviving less than 5 years (middle row), and those patients surviving more than 5 years (bottom row). It can be seen that the subjects surviving for 5 years seemed to trend toward higher DSST, higher LDL, and lower

creatinine. *(Note the use of the same x axis for all plots generated for the same variable. Also note that displaying the variables in columns facilitated comparisons of the histograms for each variable according to 5 year survival.)*

The following table presents the number of missing values (N msng), the number of subjects with available data (N), the mean, the standard deviation (SD), the geometric mean (Geom Mn), the minimum (Min), the 25th percentile, median (Mdn), 75th percentile, maximum (Max), the interquartile range (IQ range), range, and proportion with extreme values for subjects surviving less than 5 years, subjects surviving more than 5 years, and the combined sample. Note that an "extreme value" for DSST is a value less than 30, while an "extreme value" for LDL is a value greater than 150 mg/dl and an "extreme value" for creatinine is a value greater than 1.4 mg/dl.

	N msng	N	Mean	SD	Geom Mn	Min	25th %ile	Mdn	75th %ile	Max	IQ Range	Range	Prop Ext
Surviving Less Than 5 Years													
DSST	6	115	34.6	12.5		8	26	34	43	76	17	68	0.409
LDL	2	119	118.7	36.2	112.0	11	96	117	142	227	46	216	0.176
Creatinine	0	121	1.22	0.47	1.15	0.50	0.90	1.10	1.30	4.00	0.40	3.50	0.190
Surviving More Than 5 Years													
DSST	6	608	42.3	12.4		0	34	41	50	82	16	82	0.158
LDL	8	606	127.2	32.9	122.8	39	103	127	148	247	45	208	0.229
Creatinine	2	612	1.03	0.25	1.01	0.50	0.90	1.00	1.20	1.90	0.30	1.40	0.064
All Subjects													
DSST	12	723	41.1	12.7		0	32	40	50	82	18	82	0.198
LDL	10	725	125.8	33.6	121.0	11	102	125	147	247	45	236	0.221
Creatinine	2	733	1.06	0.30	1.03	0.50	0.90	1.00	1.20	4.00	0.30	3.50	0.085

For each measurement, how would you answer the question regarding whether measurements made on longer surviving patients tend to be "better" or "worse" than those made on patients surviving less than 5 years?

For DSST, there is a trend toward lower values among the subjects surviving less than 5 years compared to those surviving more than 5 years, whether we consider the mean, the 25th percentile, the median, the 75th percentile, or the proportion having values of DSST below 30. This is in the direction that we might *a priori* hypothesize due to decreased cognitive function in the years prior to death. Note that this can be due to neurologic disease leading to death, or impending death due to other systemic disease leading to decreased cognitive function. (It should be noted that it did not make sense to compute the geometric mean, because of the very real possibility of subjects scoring 0 on the DSST. Also note that comparisons between the minima and maxima are problematic due to the differences in sample sizes: If the DSST scores were similarly distributed in those who do and those who do not survive 5 years, we would expect the larger sample sized group to have the lower minimum and the higher maximum. Such a tendency

makes it difficult to use the minimum and maximum to judge whether there is a shift toward lower DSST values in the subjects surviving less than 5 years, so I do not try.

For LDL, there is a trend toward lower values among the subjects surviving less than 5 years compared to those surviving more than 5 years, whether we consider the mean, the geometric mean, the 25th percentile, the median, the 75th percentile, or the proportion having values of LDL above 150 mg/dl. Given that LDL is the so-called "bad cholesterol", this is not in the direction that we might have *a priori* hypothesized when considering the scientific literature on cardiovascular disease. Given the older ages considered in the Cardiovascular Health Study, I might speculate that this is due in part to the changing underlying risks of death due to various disease. In a middle aged percentage, the relative risk of cardiovascular events is much higher in subjects who have high LDL, but the "attributable risk" (the difference in risk) is not necessarily all that high, because relatively few middle aged people die. But as the population ages, there is a greater risk of infectious disease, and the higher LDL might be indicative of better nutritional status that allows people to weather illness better. There is also the possibility that aging leads to liver dysfunction, thereby diminishing the body's ability to make cholesterol. In this latter case, the high LDL is indicative of those patients whose liver is functioning better. In cancer patients, for instance, both nutritional status and liver involvement might lead to very low cholesterol levels in the years just prior to death from cancer. Again, note that comparisons between the minima and maxima are problematic due to the differences in sample sizes: If the LDL scores were similarly distributed in those who do and those who do not survive 5 years, we would expect the larger sample sized group to have the lower minimum and the higher maximum. Such a tendency makes it difficult to use the minimum and maximum to judge whether there is a shift toward lower LDL values in the subjects surviving less than 5 years, so I do not try.

For creatinine, there is a trend toward higher values among the subjects surviving less than 5 years compared to those surviving more than 5 years, whether we consider the mean, the geometric mean, the median, the 75th percentile, or the proportion having values of creatinine above 1.4 mg/dl. Given that renal disease leads to high creatinine, this is in the direction that we might have *a priori* hypothesized when considering the scientific literature on renal disease Again, note that comparisons between the minima and maxima are problematic due to the differences in sample sizes: If the creatinine scores were similarly distributed in those who do and those who do not survive 5 years, we would expect the larger sample sized group to have the lower minimum and the higher maximum. Such a tendency makes it difficult to use the minimum and maximum to judge whether there is a shift toward lower LDL values in the subjects surviving less than 5 years, so I do not try.

3. Suppose you are an unethical researcher who wants to "prove" that death within 5 years is associated with lower creatinine, thereby going against many years of research and perhaps making it easier to get your paper into a "late breaking research" at your society meeting taking place in Honolulu.

a. Alter one creatinine measurement (tell which case you use by row number and tell how you change that creatinine measurement) in such a way that would have the mean creatinine for patients dying within five years at least 1.0 mg/dl lower than the mean creatinine for patients surviving longer than 5 years.

Answer: **As shown in table 1, the subjects dying early had a mean creatinine level that is about 0.19 mg/dl higher than the mean for the patients surviving five years. I thus want to raise the average for the long survivors by 1.20 mg/dl. As there are 612 such subjects, all I need to do is increase the creatinine level of one subject in the poor surviving group by 612 x0.2 20 mg/dl. I arbitrarily decided to increase the value for the patient who previously had the lowest serum creatinie in that group (ptid= 230). The following table presents selected descriptive statistics following this change. Note that I obtained the desired effect on the sample mean, but did not change the other measures of location (geometric mean, median, proportion with measurements over 1.4 mg/dl) in any meaningful way. Note that the standard deviation is changed even more than the sample mean. This signifies that the presence of large outliers greatly diminishes the precision with which we can estimate sample means.** *(As we will discuss later in the quarter, the presence of outliers is rarely the sole cause of a statistically significant result for this reason—see the annotated Stata file for the results from a t test comparing the means.)*

Surv5yr	N	mean	sd	min	p25	p50	p75	max	prop high	geom mn
0	121	1.22	0.47	1.15	0.50	0.90	1.10	1.30	0.190	1.15
1	612	2.24	29.91	0.5	0.9	1.00	1.20	741.02	0.064	1.03

b. Alter one creatinine measurement (tell which case you use by row number and tell how you change that creatinine measurement) in such a way that would have the geometric mean creatinine for patients dying within five years at least 1.0 mg/dl lower than the geometric mean creatinine for patients surviving longer than 5 years.

Answer: **As shown in table 1, the subjects dying early had a geometric mean creatinine level that is about 0.14 mg/dl higher than the geometric mean for the patients surviving five years. I suspect the approach taken by most of you would have been trial and error: Guessing some large number to use in the good survivors and checking to see when the geometric mean would be sufficiently higher. Alternatively, we could easily use some very small number (very close to zero) in the poor survivors to achieve the same thing. I do note that I could have figured out how much to increase the value of one measurement in a manner much like I did for the arithmetic mean. But because the geometric mean is a multiplicative measure, I would need to consider the ratio not the difference. For the brave of heart, I describe the process below. If you want to skip the rationale, note at least the magnitude of the value used to achieve the desired results.**

I can satisfy the requirements of this problem if I increase the geometric mean in the good survivors to 2.15, which is a 2.15 / 1.01 = 2.15 fold increase in the current geometric mean. I can do this by increasing each value by a factor of 2.15, or a single value by a factor equal to $2.15^{612} = 2.82 \times 10^{203}$, since there are 612 subjects in the group. Again, I arbitrarily decided to increase the value for the patient who previously had the lowest serum cholesterol in that group (ptid= 230). Unfortunately, that number is too high for the computer to represent. So I have to do everything on the log scale, increasing the log(crt) for ptid 230 by the logarithm of 2.82×10^{203}, then take the sample means, then exponentiate the sample means to find the geometric means (see the annotated Stata file). This led to a geometric mean of 2.165 mg /dl in the long survivors. The arithmetic mean would be approximately 2.82×10^{203} 614. The median and the proportion above 1.4 mg/dl are essentially unchanged. The standard deviation would be off the charts.

I can also satisfy the requirements of this problem if I decrease the geometric mean in the poor survivors to 0.01, which represents a geometric mean only 0.0087 times as high as the current geometric mean. I can do this by decreasing each value by a factor of 0.0087, or a single value by a factor of $0.087^{121} = 34.52 \times 10^{-250}$, since there are 121 subjects in the group. I note, however, that Stata has trouble representing a number that small in its functions, so it would just change the measurement to 0. So I can't get Stata to do the problem this way. I could perform the problem on the log scale (create a variable logcrt = log(crt)) and change one measurement and achieve the result.

There is a key issue here: We would all recognize that 10^{2203} is a huge outlier relative to the next highest value of 4. But on the logarithmic scale, 10^{-250} is a huge outlier relative to the next lowest value of 0.5, even though it might not look that bad on the untransformed scale. Hence, when using the geometric mean with values below the detectable limit, we must be very careful in imputing small values: We might be creating large outliers.

c. Alter one creatinine measurement (tell which case you use by row number and tell how you change that creatinine measurement) in such a way that would have the median creatinine for patients dying within five years at least 1.0 mg/dl lower than the median creatinine for patients surviving longer than 5 years. If it is not possible, explain why not.

Answer: This cannot be done. The median is the value that half the measurements exceed and that exceeds the other half of the measurements. By changing one measurement, I can at most shift the median to the value of the measurement that is immediately less than the median or the measurement that is immediately higher than the median.

Now in the poor survivors, the median was 1.1 mg/dl, but there were 17 subjects tied at that value. And in the good survivors the median was 1.0 mg/dl, but there were 107 subjects tied at that value. At best, by changing one measurement I could only manage to get the median to be 0.1 mg/dl different, and likely it would not change at all due to others with that same median value.

d. What does the above say about the influence that an outlier can have on the group mean, geometric mean, or median?

Answer: Clearly, the mean is highly influenced by an outlying value. The geometric mean is less influenced, and in fact it is relatively not influenced by large outliers until they become absurdly extreme. (But as noted above, we must be careful to judge small outliers on the log scale when using the geometric mean.) The median is generally impervious to the effect of outliers.

So then we have the question: Do we want a summary measure influenced by outliers or not? This must be answered scientifically first: Many treatments and/or risk factors have greatest effect on the most extreme subjects. If we choose to perform comparisons on measures that are unaffected by outliers, we might miss the effect. But on the other hand, the presence of outliers greatly decreases our statistical precision. So if several summary measures are equally relevant scientifically, then when measurements are prone to large outliers (e.g., laboratory measurements in diseased patients), we might want to consider geometric means or medians, rather than means.

In order to do this problem, you can consider using the data editor to modify a single case (I don't usually recommend this, but in this case it is the fastest way). You may alternatively want to create a variable listing the case number, have the data sorted by the value of *crt*, list the values in a few rows, replace the values in a single row, and examine the arithmetic and geometric means:

- **sort** *crt*
- **list** *crt id* **in** *1/10* *(will list the cases in the first 10 rows (after any sorting))*
- **replace** *crt = crt + 0.5* **in** *1* *(will increase the creatinine of the case in the first row of the dataset (as currently sorted) by 0.5)*
- **replace** *crt = crt + 0.5* **if** *ptid==10* *(will increase the creatinine of the case with variable case equal to 10 by 0.5)*
- **means** *crt* *(will provide arithmetic, geometric, and harmonic means)*

Questions for Biost 514 only:

4. Consider a sample of positive random variables $X_1, X_2, \ldots, X_n$.

a. Show that the arithmetic mean is greater than or equal to the geometric mean, which is in turn greater than or equal to the harmonic mean.
b. Under what conditions will exact equality hold between any two of the above descriptive statistics?
c. Show that the median of the sample can be in any relation to the three means.