Biost 517: Applied Biostatistics I
Emerson, Fall 2010

Homework #2 Key
October 31, 2010

Written problems: To be handed in at the beginning of class on Wednesday, October 13, 2010.

Questions for Biost 514 and Biost 517:

1. The class web pages contain descriptions of two datasets
    - FEV and Smoking in children (fev.doc)
    - Clinical trial of idarubicin in AML (leukemia.doc)

    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:

The FEV dataset is concerned with comparing the distribution of FEV across populations that differ in their smoking habits: a statistical question of the fourth kind in my classification.

The clinical trial dataset is concerned with comparing the distribution of complete remissions (primary endpoint) and survival time (secondary endpoint) across populations that differ in the use of idarubicin in the treatment of their leukemia: a statistical question of the fourth kind in my classification.

    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:

The FEV dataset would have

- **Outcome variable: FEV**
- **Predictor of interest: the indicator of being a smoker**
- **Potential confounders: age, height, sex**
- **Potential precison variables: If the potential confounders do not turn out to be confounders, we might be interested in them as precision variables. (Looking ahead, after adjusting for confounding by age, height is definitely a precision variable.)**
- **Irrelevant: sequence number and subject ID (though the latter might be used to ensure that all subjects had only one measurement.**

The clinical trial dataset would have

- **Outcome variables: complete remission (primary), courses of chemotherapy to CR (secondary), survival time (secondary, as recorded in *fudate, onstud, status*)**
- **Predictor of interest: treatment arm**
- **Potential confounders: None (as a randomized clinical trial, we can state that there will be no confounding on average)**
- **Potential precision variables: sex, age, FAB classification, Karnofsky score, baseline hematologic profile (WBC, platelets, hemoglobin)**
- **Surrogates for the response: indicator of bone marrow transplantation and the date it occurred (note that you generally have to be alive and in remission in order to get a bone marrow transplant, so this will tend to just reflect the outcome variables, but will be far more imprecise indicators of treatment success)**
- **Irrelvant: patient ID, indicator of evaluable (we should use an intent to treat analysis), and indicator of inclusion in the interim analysis (this is of interest only as an educational exercise for sequential monitoring)**

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:

The FEV dataset has

- **Continuous variables: FEV, age, height (quantitative, ordered, ratio)**
- **Binary variables: sex, smoke (discrete, ordered or unordered—it is moot with only two values)**
- **Unordered nominal: seqnbr, subjid**

The clinical trial dataset has

- **Continuous variables: age, karn (but measured quite discretely), wbc, plt, hgb (quantitative, ordered, ratio)**
- **Nominal coding of a variable that is truly continuous: onstudy, crdate, fudate, bmtxdate (when appropriately coded, these are continuous, interval variables—there is no absolute "day 0" that would allow us to talk about a date that was double another date)**
- **Binary variables: tx, sex, eval, cr, status, bmtx, incl (discrete, ordered or unordered—it is moot)**
- **unordered nominal: karn, ptid**

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 inflammatory markers in patient survival. The data (inflamm.txt) and documentation (inflamm.doc) can be found on the class web pages. The file inflamm.txt can be downloaded and read into Stata using the command (typed all on one line)

```
infile id site age male bkrace smoker estrogen prevdis diab2 bmi systBP aai cholest crp fib ttodth death
cvddth using  inflamm.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 systolic blood pressure (SBP), the ankle:arm index (AAI), and the body mass index (BMI) at the requested levels. There are many ways to do this. One way is as follows:

- **g**enerate *hiSBP = 1*
- **replace** *hiSBP* = 0 if *systBP < 160*

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 *systBP* to also have missing for *hiSBP* (a period by itself is used as the code for missing data):

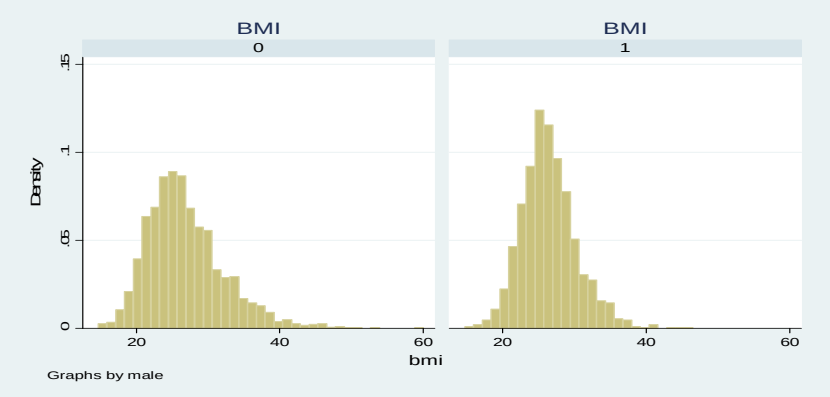
- **replace** *hiSBP* = . if *systBP* == .

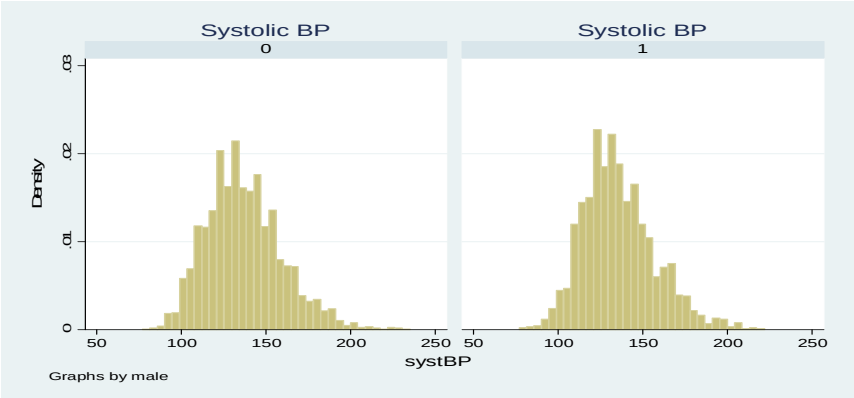
Similar variable could be created to indicate low ankle:arm index or high body mass index (BMI). (For this homework we define "high" SBP as greater than or equal to 160 mm Hg, low levels of ankle:arm index as less than or equal to 0.90, and "high" BMI as greater than 30).

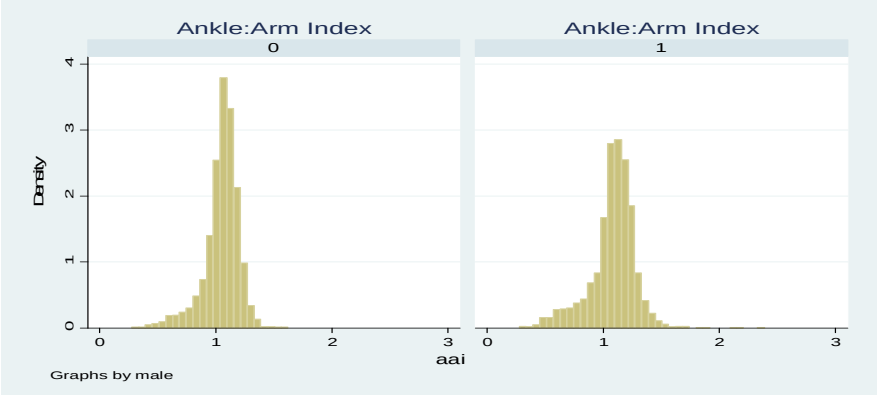
a. Using the three measurements of BMI, systolic blood pressure, and ankle:arm index, generate the following descriptive statistics for each group defined by subject sex:
   - Histogram
   - Number of cases with missing data
   - Mean
   - Geometric mean
   - Median
   - Mode (it suffices to take an 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" SBP, "high" BMI, or "low" AAI (as defined above)

Answer:

The following figures display histograms for body mass index (BMI- top row), systolic blood pressure (SBP – middle row), and ankle:arm index (AAI – bottom row) for females (left column labeled by male==0) and males (right column labeled by male==1). From these plots we see that the BMI for females is more spread out than that for males, while the AAI for males is more spread out than that for females. There also appears to be slightly skewed distributions for all of the variables.







The following tables present descriptive statistics as requested for the problem. In order to deal with space constraints, I present measures of location in the first table and measures of spread in the second table.

	N	Mean	Median	Prob Hi	Prob Lo	Geom Mn	Mode
	Females						
BMI	2,895	26.86	26.10	0.228		26.37	25.06
SBP	2,897	137.08	135.00	0.153		135.33	130.61
AAI	2,813	1.05	1.07		0.120	1.04	1.07
	Males						
BMI	2,092	26.41	26.10	0.149		26.14	25.36
SBP	2,093	135.82	133.00	0.142		134.22	129.39
AAI	2,066	1.08	1.11		0.144	1.06	1.13

	N	Std Dev	Variance	Min	25th %ile	75th %ile	Max	Range	IQR
	Females								
BMI	2,895	5.31	28.20	14.70	23.20	29.60	58.80	44.10	6.40
SBP	2,897	22.28	496.20	77.00	122.00	151.00	235.00	158.00	29.00
AAI	2,813	0.15	0.02	0.30	0.99	1.14	1.60	1.30	0.15
	Males								
BMI	2,092	3.78	14.31	15.60	23.90	28.50	46.20	30.60	4.60
SBP	2,093	21.26	451.84	79.00	121.00	148.00	219.00	140.00	27.00
AAI	2,066	0.20	0.04	0.28	1.01	1.20	2.38	2.11	0.19

In examining the summary statistics, the means, medians, geometric means, and modes are all relatively close to each other. There do not appear to be any "pathologic" outliers (i.e., outliers that would be problematic when analyzing the data), though the maximum values for the BMI, SBP, and AAI all appear further from the median than do the minimum values, thus suggesting some skewness.

For each measurement, how would you answer the question regarding whether measurements made on males tend to be "better" or "worse" than those made on females?

Answer:

Generally, higher BMI, higher SBP, and lower AAI are thought to be associated with worse cardiovascular outcomes. Though none of the differences are probably of clinical significance, whether considering the mean, median, geometric mean, mode, or probability of exceeding the specified thresholds, females tend to have very slightly "worse" BMI and SBP than men. For AAI, the differences between the sexes are again not of clinical significance, but women have the lower (and therefore "worse") mean, median, geometric mean, and mode. However, a higher proportion of men are below the specified threshold of 0.9, which is sometimes used to screen for peripheral arterial disease.

3. Suppose you are an unethical researcher who wants to "prove" that males tend to have higher SBP, thereby going against the observed data.

a. Alter one SBP measurement (tell which case you use by row number and tell how you change that SBP measurement) in such a way that would have the mean SBP for males at least 3 mm Hg higher than the mean SBP for females.

Answer:

The patient with ID 1252 was the male with the lowest SBP (79 mm Hg). So I chose to increase this value. In the real data, the mean for the males was 1.26 mm Hg less than that for the females, so if I need to increase the mean by 4.26 mm Hg. There are 2,093 males, so I need to increase the one measurement by at least 2,093 x 4.26 mm Hg. Thus if I increase the Value for ID 1252 by 2,093 x 5 mm Hg, that will more than meet the criterion: The males' mean SBP will then be 140.8 mm Hg versus 137.1 mm Hg for the females.

b. Alter one SBP measurement (tell which case you use by row number and tell how you change that SBP measurement) in such a way that would have the

geometric mean SBP for males at least 3 mm Hg higher than the geometric mean SBP for females.

Answer:

The patient with ID 1252 was the male with the lowest SBP (79 mm Hg). So I chose to increase this value. Geometric means are a multiplicative quantity. In the real data, the the geometric mean for the males was only 134.22 / 135.33 = 0.9918 that for the females. We want the geometric mean to be at least 138.33 / 135.33 = 1.02217 that for the females. This could be effected if I increased each male's measurement by a factor of 138.33 / 134.22 = 1.03062. However, since I am only allowed to change a single value, I would need to change that single value by a factor of $1.03062^{2093} = 2.609 \times 10^{27}$. Because ID 1252 had previously had a value of 79, I need to have that value at least as large as $79 \times 2.609 \times 10^{27} = 2.06 \times 10^{29}$. Before figuring all of this out, I just tried to increase the value to 10^{35} (which is represented in Stata by 1e35). The resulting analysis showed the geometric mean for males to be 139.2 mm Hg versus the females' 135.33 mm Hg. It is worth noting that the mean for males is now 4.78×10^{31} mm Hg.

c. Alter one SBP measurement (tell which case you use by row number and tell how you change that SBP measurement) in such a way that would have the median SBP for males at least 3 mm Hg higher than the median SBP for females. If it is not possible, explain why not.

Answer:

The median value for males was 133 mm Hg. There were, however, 34 males with that exact same value. So shifting the lowest measurement (ID 1252) to be the highest measurement had no impact on the median.

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

Answer:

Outliers will in general affect the mean more than the geometric mean, and they will have virtually no impact on the median. This means that if the scientific question should not be influenced by outliers, we would not want to use the mean. On the other hand, many times our "treatments" are designed to have greatest impact on the "tails" of a distribution, so use of the median in those settings will "miss the boat". The geometric mean is often viewed as a happy middle ground. It does not entirely ignore the outliers, but it downweights their influence. There are also other mechanistic reasons that we might use a geometric mean in some settings.

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 *systBP*, list the values in a few rows, replace the values in a single row, and examine the arithmetic and geometric means:

- **sort** *systBP*
- **list** *systBP id* **in** *1/10* *(will list the cases in the first 10 rows (after any sorting))*
- **replace** *systBP = systBP + 0.5* **in** *1* *(will increase the SBP of the case in the first row of the dataset (as currently sorted) by 0.5)*

- **replace** *systBP = systBP + 0.5* **if** *id==10 (will increase the SBP of the case with variable id equal to 10 by 0.5)*
- **means** *systBP (will provide arithmetic, geometric, and harmonic means)*