POLI 300 PROBLEM SET #6 11/03/10

CENTRAL TENDENCY: ANSWERS & DISCUSSION

1. (a) *modal* income: **$20,000** (income level under highest point on the frequency curve)

   (b) *median* income: approximately **$25-30,000** (income level such that vertical line from the value to the curve divides area under the curve into equal halves)

   (c) *mean* income: even more approximately **$28-34,000** ("balance point"/"center of gravity" of income level)

   In any event, for such a skewed distribution: **mode < median < mean**

2. (a) *mode*: **NO** (depends on the shape of the frequency distribution)

   (b) *median*: **NO** (depends on the shape of the frequency distribution)

   (c) *mean*: **2.5** children per household = 50 children / 20 households (regardless of the shape of the frequency distribution)

   You should be able to persuade yourself, by devising examples (also see #4 below), that 50 children can be distributed among 20 households in all sorts of different ways with all sorts of different modes and medians (and standard deviations). But, by definition, the mean of every such distribution is 50/20 = **2.5**.

3. (a) *mode*: **1 child** Most frequently occurring value of the variable NUMBER OF CHILDREN

   (b) *median*: **2 children** No more than half of households have fewer and no more that half have more

   Note. The median is *not* 5, i.e., not the *midpoint* of the range — 76% of households have fewer than 5 kids. Lots of people made this mistake! If you propose some value as the median, always check to make sure no more than half the cases have higher values and no more than half lower values.

   (c) *mean*: **2.96 children**

   mean $= (0\times.1)+(1\times.26)+(2\times.16)+(3\times.14)+(4\times.1)+(5\times.08)+(6\times.06)+$
   $(7\times.04)+(8\times.03)+(9\times.02)+(10\times.01)$
   $= 0 +.26 +.32 +.42 +.4 +.4 +.36 +.28 +.24 +.18 +.1 =$ **2.96**

4. (A) **TRUE** (B) **TRUE** (C) **FALSE**

   This is similar to question as #2. The mean of the distribution is $100/*n*, however the money is divided up. But all other statistics depend on *how* the money is divided up. For example, suppose n = 10 and the $100 is initially divided equally and then redistributed so that one person has $100 and everyone else has nothing.

	Before	*After*
Mode	$10	$0
Median	$10	$0
Mean	$10	$10

5. The maximum values of IDEOLOGY and HEALTH INSURANCE is 5; the maximum values of ABORTION is 4. Quite a few students gave "averages" that were (much) higher than these maximums, which they should have recognized couldn't possibly be correct.. For each mode, find the *value* of the variables with the *greatest* (absolute or relative) *frequency*; note that student HEALTH INSURANCE has two modes (tied for the greatest frequency). For each median, find the values at which the cumulative relative frequency first exceeds 50%. For student IDEOLOGY, note that "Liberal" plus "Sightly Liberal" includes 25/50 = 50% of the cases, and "Moderate" plus "Slightly Conservative" plus "Conservative" also includes 25/50 = 50% of the cases, so by convention the mode is the midpoint between the two observed values right in the middle of the ranked data, i.e., 2.5 (between "Slightly Liberal" and "Moderate").

	IDEOLOGY		**HEALTH INSURANCE**		**ABORTION**	
	ANES	Students	ANES	Students	ANES	Students
Modal Value	3	1	1	2&4	4	4
Median Value	3	2.5	3	3	3	4
Mean Value	3.24	2.66	2.72	2.8	2.83	3.32

The (whole number) modal and median values can also be reported in terms of the *value labels* (i.e., "Liberal," "Slightly Liberal," "Mostly Government," "Never Permit," etc.) instead of the code values. But the mean needs to be calculated numerically and may have fractional values.

Given the frequency distributions, there are the two ways to calculate the mean. Here are the calculations for the mean value of ANES IDEOLOGY and the others can be calculated in like manner.

Value × Absolute Frequency	Value × Adjusted Relative Frequency
1 × 291 = 291	1 × .167 = .167
2 × 205 = 410	2 × .118 = .236
3 × 506 = 1518	3 × .291 = .873
4 × 278 = 1112	4 × .160 = .640
5 × 461 = 2305	5 × .265 = 1.325
5636/1740 = **3.239**	**3.241**

Since SPSS rounds percentages off to the nearest one-tenth of one percent, there is some rounding error in the calculation based on relative frequencies, which accounts for the slight difference in the two answers. (The first calculation is exactly right, until we round off the final answer: 5636/1740 = 3.23908046....)

6. The chart depicts snowfall over 32 seasons. In 9 cases, snowfall is "above normal", in 20 it is "below normal," and in 3 it is (almost) exactly normal. Because twice as many seasons are below normal as above normal, it certainly appears that the "normal" snowfall of 15.2 inches is *not* the median snowfall, which instead is several inches below 15.2. We can determine the median snowfall by lowering the "normal" line until no more than half the bars extend above the line and no more than half fall short of the line. (It appears that 2003-4 and 2004-5 are the middle pair of cases, giving a median of about 12".) We can also check whether 15.2 inches is the mean snowfall by adding up all the deviations from 15.2 and seeing whether they add up to about zero. Note that most seasons have modest snowfalls, while some have major snowfalls which pull the mean above the median. Thus most winters have "below average" (i.e., mean) snowfalls.

   Note. Some of the points above implicitly assume that this (non-random) sample of 32 seasons is more or less representative of all the seasons going back to the beginning of record keeping.

7. (a) All the average scores will *change*, but they do not have to be *recalculated* from scratch (rather just be adjusted). According to the original calculations, the most frequent score (i.e., the mode) was 37 and the score at the middle of the ordered list of scores (i.e., the median) was 35. Every score, including the modal and median scores, is now 3 points higher, so the **new mode** is 37 + 3 = **40** and the **new median** is 35 + 3 = **38**. The original sum of scores was $n \times 32.238$. The new sum of scores is $(n \times 32.238) + (n \times 3)$, so the **new mean** is 32.238 + 3 = **35.238**. In general, if any positive or negative number is added to all the of scores, each average changes by the same amount.

   (b) Likewise, if every score is multiplied (e.g., doubled) or divided by some number, each average is changed by the same factor. So (including the 3 point adjustment discussed above) the **new mode** is new mode is 40 × 2 = **80**, the **new median** is 38 × 2 = **76**, and the **new mean** is 35.238 × 2 = **70.476** (or **74**, **70**, and **64.476** excluding the prior 3 point adjustment).

   (c) Only the top 20% of scores are changed, while (by definition) almost 50% of the scores lie above the median score, so the median score is not affected. On the other hand, the sum of all scores is increased by $.2n \times 5$, so the mean is increased by $(.2n \times 5)/\ n =$ **1** point above what it otherwise would be. (Likewise, "if the rich get richer and everyone else stays the same," mean [per capita] income increases, while median income stays the same.) Note also that the mean goes up one point and the median stays the same if 5 points are added to the scores of the bottom 20% of the class. ("If the poor get a bit richer and everyone else stays the same," mean [per capita] income increases, while median income stays the same.)

8. As a practical matter, the first question comes down this: which variable (HEIGHT or WEIGHT) has a frequency distribution that is closest to *symmetric* (in which case the median and median are just about the same) and which is more clearly *skewed* (in which case the mean is gets pulled in the direction of the long thin tail). It should be clear that the frequency distribution for WEIGHT is *not symmetric*. The average (mean) weight of an American adult is probably about 165 pounds. As a empirical matter, no (or hardly any) negative deviations from 165 exceed about −80 (that is, hardly any American adult weighs less than about 85 pounds) and, at the logical extreme, certainly no negative deviations exceed −165 (that is, no one can weight less than zero pounds). On the other hand, quite a lot of positive deviations exceed +80 (that is, quite a lot of American adults weight more than 245 pounds) and a few exceed +165 (that is, a few American adults weigh more than 330 pounds). The average height of American adults is probably about 5'8" and (going through the same kind of thinking as above for WEIGHT) it seems clear that positive and negative deviations from the average approximately balance, so the distribution is approximately symmetric. Likewise most people these day have 0, or 1, or 2 siblings. But some people have many more siblings, while nobody can have fewer than zero. So this distribution is also somewhat skewed with a long tail in the high direction. In this respect NUMBER OF SIBLINGS resembles WEIGHT (and INCOME), though it is discrete while the other two are continuous.

Many (indeed most) students correctly said that the median and mean heights would be about the same, while mean weight would be greater than median weight, but they attributed this to the fact that there are “smaller differences” (i.e., less *dispersion*) among people in with respect to height than weight. While this sentence is true in terms one summary statistic, it is meaningless in terms of the standard deviation (or range), because we would be comparing the proverbial “apples and oranges.” The SD and range are in the same units are the variables in question, and HEIGHT and WEIGHT are measure in different and non-comparable units. Specifically the SD of HEIGHT among American adults may be about 3 inches and the SD of WEIGHT may be about 35 pounds and, while the number 25 is certainly greater than number 3, we cannot say that an SD of **25 pounds** is greater than an SD of **3 inches**, because (for example) we could just as well express the latter SD as **0.0125 tons**, and 0.0125 is a smaller number than 3. However, the commonsense perception that WEIGHT “varies more” than HEIGHT is confirmed if we measure dispersion by the *ratio measure of dispersion* called the *coefficient of variation* (see Handout #7 and corresponding PPTs and also just below). But none of this is relevant to the question concerning the median and mean values of HEIGHT and WEIGHT, which depends on the *symmetry* vs. *skewness* of the distributions of HEIGHT and WEIGHT values, not the *magnitude* of their *dispersions*.

The *coefficient of variation* measures the *amount of dispersion in a variable relative to its mean value*. Consider the distribution of heights in the adult American population. Suppose (as we did above) that the average (mean) height of American adults is about 68" (5'8"). However, (almost) no one is exactly 68" tall. Let’s “guesstimate” the average (mean) height of all American who are less than 68" tall. Plausibly this is about 65", so the average negative deviation from the mean is about 3". The average positive deviation is probably about 3" also, so the overall average (absolute) deviation from the mean is about 3", compared with a mean height of 68". Thus the coefficient of variation is something like 3"/68" = .044. In the same way, consider the distribution of weights in the adult American population. Suppose (as we did above) that the average (mean) weight of American adults is about 165 pounds. However, (almost) no one weighs exactly 165 pounds. Let's "guesstimate" the average (mean) weight of all American who weigh less than 165. Plausibly this is about 135, so the average negative deviation from the mean is about 30 pounds. For the reasons noted above, the average positive deviation is certainly more than this and is maybe something like 45 pounds — so the overall average (absolute) deviation from the mean is something like 35-40 pounds, compared with a mean weight of about 165 pounds. So the coefficient of variation is something like 37.5 pounds/165 pounds = .23, or about five time the similar ratio for height.

Note that, since the coefficient of variation is the ratio of two numbers measured in the same units (e.g., inches, pounds, etc.), the units cancel out and its value is a *pure number* that is independent of the particular units of measurement used. For example, if we "went metric" and measured everyone's heights and weights in meters and kilograms rather than inches and pounds, the numerical values of the mean and SD of both variables would change in like manner but the coefficients of variation would remain the same.

9. Total payroll is $480,000 and is divided eight ways. So the mean of the eight salaries is **$60,000**. But all seven employees other than the owner earn [much] less than this "average." The median and mode are both **$22,000**.

   Note. A few students said the modal salary is $270,000. The modal value is the *most frequently* occurring value of the variable, not the *biggest* value.

10. Negatives deviations from the average house price certainly cannot exceed $159K (houses can't have negative prices) but large mansions certainly have positive deviations of millions of dollars. So the distribution of house prices certainly is not symmetric and instead is skewed (like income) with a long thin in the direction of high prices. In such a skewed distribution, the mean is greater than the median, so **$129,900 = median** and **$159,000 = mean**.

    Note. On this question and several others, it was quite common for students to say that the mean is "generally" greater than the median, but this is not "generally" true. If the distribution is symmetric, the two averages are the same. If the distribution is asymmetric the mean is pulled in the direction of the long thin tail, regardless of whether that tail points in the low or high direction. (In the NUMBER OF PROBLEM SETS example, the tail was in the low direction, so the mean was lower than the median.)

11. The "total age" (Σx) of all people originally in the room is 5 × 30 = 150, since the mean = $\Sigma x / n$). Adding the new person, "total age" becomes 150 + 36 = 186, so the new mean age is 186/6 = **31**.

    You cannot determine the new median age without knowing more about the actual frequency distribution of ages. (Recall Problems #2 and #4 above.) However, the median age certainly has not gone down and most likely has gone up. In fact we can deduce that *the new median age cannot be higher than 33*. Suppose we rank the five people originally in the room into *ascending order* by age. We know the age of the middle (third) person on the list is 30. When the 36 year-old person is added to the list, the new median value is the midpoint between the ages of the third and fourth persons (i.e., the middle pair) on the expanded list. If the 36 year-old happens to be fourth person on the expanded list, the new median age is 33 (midpoint between 36 and 30); otherwise (i.e., if there is someone whose age is between and 30 and 36) the new median age is less than 33 (but not less than 30). (If the fourth person is also 30, the median is unchanged.)

12. You are traveling faster than half of the other vehicles and slower than half. So you are traveling at the **median** speed.

13. Except for one logically possible but highly unlikely possibility, the $2.36 million salary must be the **mean** salary. If $2.36 million were the median NBA player salary (and if no players have identical salaries), by definition 205 (rather than 139) players would have higher salaries (and the other 205 lower). ($2.36 million could be the median salary *only if* at least 67 players had *exactly* the same salary of $2,360,000.00.)

    Note. To say $2.36 million is the mean just because the news article calls it "the average" is inadequate. The mode, median, and mean (and geometric mean and harmonic mean) are all different kinds of "averages." This observation also applies to some answers for several other questions.

16. If the genie tells you what the mean will, you know *exactly* how many cans of soda you need. In particular, if the genie tells you that the mean will be 5, you know that you need exactly 30 × 5 = **150 cans of soda** — you won't run out and you'll have none left over. If the genie tells you that the median will be 3, you know by logical deduction that you need **at least 48 cans**. (You need exactly 48 cans in the event that 16 guests drink 3 cans each and the other 14 drink none.) But you also know as a practical matter that you need a good more than this (because undoubtedly some guests will drink more than 3 cans and some guests will drink 1 or 2 cans), but you cannot tell from the median value of 3 how many more of each there will be.