DISPERSION: ANSWERS & DISCUSSION

1. (a) The income level of the household at the 90th percentile (10% have higher income) appears to be (very) approximately $75,000 (see PS #5C, question 5(f)). The income level of the household at the 10th percentile (10% have lower income) appears to be approximately $10,000 (see PS #5C, question 5(e)). So: **Interdecile range** ≈ $75,000 - $10,000 = **$65,000**. The income level of the household at the 75th percentile (25% have higher income) appears to be (very) approximately $45,000. The income level of the household at the 25th percentile (25% have lower income) appears to be approximately $18,000. So: **Interquartile range** ≈ $45,000 - $18,000 = **$27,000**. *Note*: each range has a single dollar value, i.e., range is the *difference* (or "distance" or "interval") between the high and low values, not the high and low values themselves.

   (b) The simple range [maximum value – minimum value] is not usable here because of the "open-ended" (at the upper end) nature of the income variable [so we don't know the maximum value].

   (c) Mean income ("center of gravity" or "balance point" of the distribution) appears to be about $28-34,000 — let's say the mean is $30,000. Cut off the part of the curve that is above the mean and guess where the portion of the curve that is below the mean would "balance" — maybe about $18,000. Thus the average deviation from the mean among all households that are below the mean is about $18,00 – $30,000 or –$12,000. Now look at the part of the curve that is above the mean and guess where it would "balance"(remembering that a few cases [Bill Gates, etc.] are way off the graph) — let's say about $60,000. Thus the average deviation from the mean among all households that are above the mean is about +$30,000. So the average absolute deviation from the mean (the **MD**) among all households is approximately the average of $12,000 and $30,000, or about **$21,000**. The **standard deviation** is somewhat larger than the MD (Bill Gates and all those CEOs and rock and sports starts have very large positive deviations that are even more enormous, relative to typical deviations, after they are squared) — perhaps about **$26-30,000**. [Some students said SD = $7,500 with no explanation — possibly because $7.5 is the square root of the midpoint of the income scale shown in the diagram?] Note that these estimates imply that, for this data, the SD is approximately equal to the interquartile range. In a *normal* distribution, the interquartile range is somewhat larger than the SD (it is equal to about 1.35 × SD; see Figure 1 in Handout #8). In this skewed distribution the relatively few cases with very high values increase the SD but have no impact on the interquartile range. (If the richest 24% of households got a lot richer, this would increase the SD [and the mean] but would have no effect on the interquartile range [or the median].)

   (d) If the income distribution were more equal, the frequency diagram would be more "bunched up" near its mean value, and its SD [and ranges] would be smaller.

(e) If the income distribution were more unequal, the frequency diagram would be even more "spread out" with area cut out of the near-average portion and redistributed to both the rich and poor extremes (the rich and poor would be gaining *cases*, not income, of course), and its SD [and ranges] would be larger. ***Note***: many students talked about "the frequency distribution increasing/decreasing"; this really does not mean anything — the area under the curve always represents 100% of the cases.

(f) If everyone's income doubled, the frequency curve would remain anchored at the $ = 0 point but it would be "stretched" upwards to twice its present "width." (Alternatively, and more simply, we could keep the curve as it is but rescale the horizontal intervals by doubling every value.) All averages, ranges, and the SD are doubled (and the variance is "doubled squared," i.e., increased fourfold.)

(g) If everyone's income were increased by a constant amount, e.g., $25,000, the shape of the curve would remain just same but the whole curve would "slide" to the right by that amount. All averages would increase by the same amount. Ranges and the SD would be unchanged.

***Note*.** While the SD and other (interval) measures of dispersion double in (f) and stay the same in (g), income *inequality* (in the ordinary understanding of the term) is greatly reduced by giving everyone $25,000 in the manner of (g). This is because the SD takes account only of the *interval* property INCOME, while our ordinary understanding of income inequality takes account of the *ratio* property of INCOME. Income inequality as measured by the *coefficient of variation* and the *Gini Index*, both of which take account of the ratio property, is *unchanged* in (f) and substantially *reduced* in the (g). See the discussion of *Dispersion in Ratio Variables* in the handout on *Measures of Dispersion* and also Problems #8, #11, and #12 below.

2. **NO**, you know nothing about dispersion at all. In general, the smallest possible SD is zero (which occurs whenever every case has exactly the same value — necessarily the mean [and median and modal] value). But in this case, we know that *no* household can have the mean value of 2.5 children, because the variable is *discrete* and has only whole-number values. Every household is as close to the mean of 2.5 as possible when half have 2 children and the other half have 3 children, in which case the **SD** (and mean deviation) = **0.5** . (***Note***: some students answered SD = 1; that would be the minimum *range*. The maximum SD [of about 10.9] occur would occur in the [highly unlikely] event that one household has all 50 children.)

3. Guess at SD. The mean is about 3 children (see A&D for PS #6, Question 3). 16% of the cases deviate downward by 1 one child, 26% by 2, 10% by 3, so the average negative deviation is somewhat less than 2 children. Most upward deviations are only 1 or 2 children but a small proportion are considerably greater, so the average positive deviation is somewhat more than 2 children. 14% of the cases have (almost) zero deviation. So the *mean deviation* is probably a bit under 2 children. The *standard deviation* is almost always somewhat (about 20-50%) larger (and never smaller) than the mean deviation, so the SD is probably

somewhere between 2.2 and 2.5 children. [The calculated SD = 2.4 children; see below. The *variance* is 5.76 children squared (whatever that means).]

Since this frequency distribution looks something like the income distribution in Problem #1, we might expect the SD and interquartile range to be approximately equal, as in the income case. This comparison is hindered by the fact that NUMBER OF CHILDREN is a *discrete* variable while INCOME is *continuous*. 10% of households have no children and 36% (cumulating downward) have 1 or fewer children and just under a quarter (24% cumulating upwards) have 5 or more children. Thus the value of the case at the 25th percentile is 1 and the value of the case at the 75th percentile is 4, so the interquartile range is about 3 — a bit larger than our estimate for the SD.

For the record, here is how the SD may be calculated. Remember the mean of the data was previously (in PS #6) calculated to be 2.96. It is interesting to see what happens when we round this off to 3, which makes the calculations much less burdensome. As you can see, such rounding has virtually no effect on the calculated SD.

	Use Mean = 2.96		*Use Mean* = 3	
Value	*Dev.*	$(Dev.)^2 \times Freq.$	*Dev.*	$(Dev.)^2 \times Freq.$
0	−2.96	$8.76 \times .10 = .876$	−3	$9 \times .10 = .90$
1	−1.96	$3.84 \times .26 = .998$	−2	$4 \times .26 = 1.04$
2	−0.96	$0.92 \times .16 = .147$	−1	$1 \times .16 = .16$
3	+0.04	$0.00 \times .14 = .000$	0	$0 \times .14 = .00$
4	+1.04	$1.08 \times .10 = .108$	+1	$1 \times .10 = .10$
5	+2.04	$4.16 \times .08 = .333$	+2	$4 \times .08 = .32$
6	+3.04	$9.24 \times .06 = .554$	+3	$9 \times .06 = .54$
7	+4.04	$16.32 \times .04 = .653$	+4	$16 \times .04 = .64$
8	+5.04	$25.40 \times .03 = .762$	+5	$25 \times .03 = .75$
9	+6.04	$36.48 \times .02 = .730$	+6	$36 \times .02 = .72$
10	+7.04	$49.56 \times .01 = .496$	+7	$49 \times .01 = .49$
	Variance	5.657		5.66
	SD	2.378		2.379

4. The standard deviation is based on (squared) deviations from the mean, and we concluded in PS #6 that "normal" snowfall is mean snowfall (over an extended period of time). So we should examine the deviations of the heights of the bars from 15.2" for each year. The four highlighted years deviate upwards by about by about 23, 16, 31, and 25 inches (excluding 2009-10, for which the data incomplete). The three other above normal years (all coming in the first five years) deviate upwards by an average amount of about 10. So the sum of the positive deviation is about 23 + 16 + 31 + 25 + 30 = 125 inches. When I stare at the charts, it appears that average snowfall in the 19 below normal years is about 10 inches, for an average negative deviation of about 5 inches, so the sum of the negative deviations is about −95 inches. (There are also three years when snowfall was almost exactly "normal.") So the

sum of the *absolute* deviations is about 120, and the *mean deviation* is about 120/29 ≈4 inches. The *standard deviation* is typically somewhat lager than the mean deviation, so the SD is about 5 to 6 inches.

5.

Case #	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$\lvert x_i - \bar{x} \rvert$
1	2	−3	9	3
2	7	+2	4	2
3	1	−4	16	4
4	15	+10	100	10
5	4	−1	1	1
6	6	+1	1	1
7	2	−3	9	3
8	3	−2	4	2
Column sum:	40	0	144	26
Column mean:	5 (= $\bar{x}$)		18 = variance	3.25 = MD
			SD = $\sqrt{18} \approx 4.24$	

Note 1. You are guaranteed to get a problem like this on the second midterm. In order to get credit, you *must show each step in the calculation*. By doing so, you can also get *partial credit even if your final answer is wrong* due to some arithmetic mistake. I strongly recommend that you show each step by following the format of standard worksheet used above (and also in the handout on How To Calculate a Standard Deviation). Note: the mean deviation was also calculated (MD = 3.25) above for comparison and illustration only — you were not asked to do this.

Note 2. A number of students who evidently had no difficulty in applying the SD formula did not perform the simple check that **sum of deviations = 0** and therefore did not notice that they had made careless errors in calculating the mean, which threw off all subsequent calculations.

6. Some students redid the calculations from scratch, but this is unnecessary.

(a) All scores go up by 5 but so does the mean, so deviations from the mean are unchanged, and the SD is unchanged.

(b) All scores (and the mean) go up fivefold, so deviations from the mean also go up fivefold, and so does the SD. (The variance goes up 25-fold.)

7. (a) The mean is always \$100/*n* (where *n* is the size of the group), regardless of how the money is distributed. So **No** — the mean cannot change.

(b) **No**

(c) **Yes**, the SD could be as little as zero, if *everyone* gets the mean of \$100/*n*. The SD (and MD) would be as great as possible when one member of the group gets the whole \$100 and the rest get nothing.

(d) **No** — for example, if John initially has \$75 and Jane has \$25 (and everyone else has \$0), and the redistribution gives Jane \$75 and John \$25 (and keeps everyone one else at \$0), the SD is unchanged.

(e) See (c) above

(f) See (c) above

8. *Median* = $152 \times 0.453 = 68.9$ kg
Mean = $159 \times 0.453 = 72.0$ kg
Range = $217 \times 0.453 = 98.3$ kg
SD = $24 \times 0.453 = 10.87$ kg
Variance = $10.87^2 = 118$ kg^2 $(= 576 \times 0.453^2)$

Note. The variance cannot be $576 \times 0.453 = 261$, because $261 \neq 10.87^2$ [many students were "tricked" by this].

9. (a) Deviations for (i) and (iii) = −10, 0, +10; squared deviations = 100, 0, 100; and Deviations for (ii) = -50, 0, +50; squared deviations = 2500, 0, 2500; so

(i) SD = square root of 200/3 or about **8.18**

(ii) SD = square root of 5000/3 or about **40.8**

(iii) SD = square root of 200/3 or about **8.18**

Many people said the SDs were 10, 50, and 10, evidently forgetting that in each distribution there is a third case with zero deviation from the mean.

Despite the different SDs, (i) and (ii) seem to be "equally unequal" — in both situations, Ann earns three times what Jim earns and Bob's earnings are just average. Despite having the same SD as (i), (iii) seems much more equal — Ann earns only about 1.18 times what Jim earns and again Bob's earning are the average.

(b) These examples show that the standard deviation (based only on the interval properties of variables) does *not* capture our sense of *inequality* with respect to *ratio variables* like earnings, income, house prices. (Recall the ***Note*** on p. 2.) First note that the "inequality comparisons" we made with respect to Ann's and Jim's earnings were indeed *ratio*, not interval, comparisons. Note also that, once we get to the second column in the SD worksheet (the deviations from the mean in each case), the calculations for (i) and (iii) are identical. The difference between (i) and (iii) is that in (i) the deviations from the mean are large in relation to the mean itself (another ratio comparison), whereas in (iii) the deviations are small relative to the (much larger) mean (another ratio comparison). A simple measure of inequality in ratio variables is the *coefficient of variation* (CV), which is equal to the *standard deviation divided by the mean*. This measure answers this question: how big on average are the deviations from the mean in each case (as measured by the SD) relative to the common baseline for all these deviations (i.e., the mean)?

(i) CV = 8.18/20 = 0.41

(ii) CV = 40.8/100 = 0.41
(iii) CV = 8.18/120 = 0.068

Here's another simple example illustrating the same general point. Suppose Jack is born on his older brother Joe's 4th birthday. The *interval difference* in their ages — a range of 4 years and an SD of 2 years (check it using the formula) — will never change over their lifetimes. (Joe's deviation from their mean age is always +2 years and Jack's is always −2 years.) But the *ratio difference* in their ages diminishes with time. In substantive terms, their age difference starts out being very major (there's a huge difference between a 4-year old and an infant), remains substantial a dozen years later (when high-schooler Joe regards middle-schooler Jack as a twerpy kid brother) but becomes trivial when they are middle-aged adults or older. What is important is the ratio, not interval, comparison of their ages, and this is what the coefficient of variation captures.

10. (a) The **range** is 5 − 1 = **4** for *all* the ideology items in the 1992 NES (with 2255 respondents).

(b) Even in a small sample like the student survey, we often find one or more cases with each extreme value, so even in such small *n* data the range would be 4. That the range is unhelpful for variables of this sort was noted in Handout #7 and in class.

Note 1. Quite a few students gave the "range of the *frequencies*" (that is, something like Bush: 50% − 7% = 43%, Clinton: 36% − 7% = 29%, Perot: 29% − 12% = 17%), rather the range of the *observed values* of the variable IDEOLOGY (from 1 to 5 in all charts).

Note 2. A related, *very common*, and *fundamental mistake* is to think that when the frequency bars are all about the same height (as in the Perot case), the SD is therefore small, and that when the frequency bars vary greatly in height (as in the Bush case) the SD is therefor large. This again confuses dispersion in *frequencies* (bar heights) with *dispersion in the observed values of cases* (along the horizontal scale).

(c) Next, let's form some expectations. Clearly we expect the *mean perception* of Clinton to be somewhat on the liberal side, the mean perception of Bush to be somewhat on the conservative side, and the mean perception of Perot to be somewhere in between. Forming expectations about the *dispersion of perceptions* of the ideological positions of the candidates is trickier but we might develop them in the following way. Perot in 1992 was a completely novel candidate, who was unconnected with either political party and whose rhetoric did not follow any predictable liberal or conservative pattern, so — even while we expect perceptions of Perot's ideological position to be *on average* centrist — we might expect a lot of dispersion in these perceptions, i.e., for them to be pretty much "all over the map." Governor Clinton, like Perot but unlike the incumbent President Bush, was a novelty to most voters (in 1992, though that's hard to believe now), and he also stressed that he was a different kind of "new Democrat," so we might expect that perceptions of Clinton would be somewhat more

dispersed than those of the very well known incumbent President (and former Vice President) Bush. Thus we might expect perceptions of the ideological positions of the candidates to be ordered (from lowest to highest) in terms of dispersion as follows: Bush (most familiar), Clinton, Perot (most novel).

To my eye, these expectations appear are pretty well supported by *visual inspection* of the graphs. They are not fully supported by the *calculated* SDs, however. Here are the actual statistics (calculated by SPSS) [plus the *mean deviations* (calculated by hand with some approximation)]:

Variable	*Mean*	*Std. Dev*	*Mean Dev.*
Bush Ideology	3.96	1.29	1.04
Clinton Ideology	2.24	1.20	1.01
Perot Ideology	3.27	1.44	1.25

First of all, our expectations with regard to means are fully confirmed. And the expectation that perceptions of Perot would be "all over the map" is reflected in Perot Ideology having the highest SD (and MD). (If all five bars were the exactly the same height, the SD would be $\sqrt{2} \approx 1.4$, and the Perot is pretty close to this.) What is counterintuitive is that Bush Ideology has the second highest SD, though its distribution appears to the eye to be the most *concentrated*. This results from the fact that the Bush distribution is also highly *skewed* in the conservative direction, so its mean is far off-center. Nevertheless, some (presumably confused) respondents put Bush at the liberal end of the scale, so they have *very large* deviations from the mean, which become even larger *squared* deviations, with the result that the SD is quite large. This reflects the fact that the SD does not work very well as a measure of dispersion when many cases are at *one end or other of a measuring scale* which has *definite upper and/or lower bounds*. (In this case, the *mean deviation* comes closer to meeting our expectations.)

There is another (much simpler) noninterval-level measure of dispersion (which can be used with ordinal, or even nominal, variables) that may better reflect our intuitive impressions concerning bar graphs of this sort. This is the *variation ratio*, defined as

$$v = 1 - \frac{n(\text{modal})}{n} = 1 - f(\text{modal})$$

where "n(modal)" is the number of cases with the modal value, "n" is (as usual) the total number of cases (excluding missing data), and "f(modal)" is the *adjusted relative frequency of the modal value*. In words, v is simply the fraction of cases that do *not* have the modal value. Put otherwise, it is inversely related to the height of the highest bar in a frequency bar chart. In this data, the magnitude of v conforms with our expectations:

v(Bush Ideology) = .502 (50.2% of cases do not have the modal value of 5)
v(Clinton Ideology) = .652 (65.2% of cases do not have the modal value of 1)
v(Perot Ideology) = .705 (70.5% of cases do not have the modal value of 5)

This is a very crude measure of dispersion, however, because it takes account of only the *nominal* ("same or different") property of the variable. Like the mode (and because it is based on the mode), its value can change erratically with small changes in the data.

(e) The minimum possible SD for *any* variable is zero; in this case, the minimum would result when *all* respondents have *identical* perceptions of the named candidate [regardless of what that perception was].

(f) The maximum possible SD for any variable measured on a "bounded" scale like this results when half the cases have one extreme value and the other half have the other extreme value. In this case, the mean is 3, the deviation from the mean for every case is either +2 or −2, the squared deviation for every case is 4, so the average squared deviation (variance) is 4, and the SD (and also the MD) is 2. So the actual SD must fall between 2 and 0, and we might expect the SD to average about 1, more or less according to the degree of consensus among the respondents.

11. (a) House seats and electoral votes have the same range: 53 − 1 = 55 − 3 = 52 (and the same interdecile and interquartile ranges).

(b/c) Each state also has the same deviation from the mean with respect to House seats and electoral votes. The mean deviation is 7.164, and the average squared deviation (variance) is 88.022, so the SD is 9.582. (*Note*: these calculations treat DC as if it were a state.)

(d) The "constant bonus" of two electoral votes based on Senate representation makes the electoral vote distribution less unfavorable to small states than it otherwise would be (i.e., less unequal). While CA has a 53 times more House seats than DE, it has only 18.33 times more electoral votes. The mean or standard deviation (being based only on the interval property of the variable) do not reflect these differences in ratios, but the coefficient of variation does:

House seats: mean 8.549; coefficient of variation: 9.582/8.549 = 1.122

Electoral votes: mean = 10.549; coefficient of variation: 9.582/10.549 = 0.908

12. (a) 5 5 5 5 (SD = 0) Any set of four *identical* numbers is equally good.

(b) 3 4 5 6 (SD = square root of 5/4 ≈1.12) Any set of four *consecutive* (and thus as "close together" as possible) numbers is equally good.

(c) 0 0 10 10 (SD = 5) Any other set of four numbers has a smaller SD.

(d) 0 1 9 10 (SD = square root of 82/4 ≈4.53) Any other set of four different numbers has a smaller SD.

13. If each teacher's salary is increased by $1000, the entire salary distribution shifts upwards by $1000), so

(a) the median and mean salaries are likewise increased.

Such an across-the-board dollar increase (or decrease) does not change *dispersion*, so

(b) the (total and) interquartile range is unchanged, and
(c) the SD (and MD) are likewise unchanged.

However, an across-the-board dollar increase (of $1000 or any other amount) reduces *inequality* in salaries (because the salaries of lower paid teachers are increased more in ratio (percentage) terms than those of higher paid teachers. In particular, the *coefficient of variation* (or *Gini Index*) will be *reduced* by the across-the-board dollar increase, since the SD remains the same while the mean is *increased* by $1000.

14. If each teacher's salary is increased by 5%, the mean and median salaries, like all salaries, go up by the same 5%. (Which average goes up more in dollar terms depends on which is higher, which in turn depends on whether salary distribution is skewed up or down — of course, if the distribution is symmetric, the mean and median are the same.) Interval measures of dispersion (except the coefficient of variation) are likewise increased by 5% (since, in dollar terms, higher salaries go up more than lower salaries). However, the degree of inequality in salaries is unchanged. In particular, the *coefficient of variation* (and the *Gini Index*)will be *unchanged* by the 5% increase; both the SD and the mean are increased by the same 5%, so their ratio remains the same.

15. Ty Cobb's 1911 batting average as standard score: +4.15
Ted Williams' 1941 batting average as a standard score: +4.26
George Brett's 1980 batting average as a standard score: +4.07

Williams (widely called the best "pure hitter" ever when he died some years ago) was apparently the most (literally) "outstanding" batter. Williams .406 average was more outstanding than Cobb's .420 because of the quite striking reduction in the *dispersion* of batting averages from the 1910s to the 1940s. Brett's .390 average put him into almost the same league as Cobb and Williams in part because of *lower mean* batting averages in the 1970s but more because of the further *decline in dispersion*.

Note that in the Cobb era, one standard deviation above the mean was .303, which implies (assuming, as we are told, that the distributions are reasonably normal) that more than 16% of batters had batting averages above .300. In contrast, in the most recent decade .300 is 1.23 standard deviations above the mean, implying that only about 11% of batters have batting averages higher than .300.

OVER ==>

16. (a) 75 inches is two standard deviations above the mean, so by the 68-95-99.7% rule, about **2.5%** are taller than this (and about 2.5% are shorter than 65 inches, and about 95% are between 65 and 75 inches).

(b) As noted above, between **65 and 75** inches.

(c) 67.5 inches is one SD below the mean, so about **16%** are shorter than this (and about 16% are taller than 72.5, and about 68% are between 67.5 and 72.5 inches).

A note on the logically maximum SD.

(a) *If the range of data is fixed* (like in the liberal-conservative bar graphs), the maximum SD (and MD) occurs when half the cases have the minimum value and half the maximum ("maximum polarization"), in which case MD = SD = Range/2

(b) *If the variable is ratio and mean/total of the values is fixed* (like dividing up a fixed sum of money), the maximum SD occurs when one case get the entire sum and everyone else get nothing (a single "outlier").