Biost 517: Applied Biostatistics I
Emerson, Fall 2011

Homework #3 Key
October 12, 2011

Written problems: To be handed in at the beginning of class on Wednesday, October 19, 2011.

*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.)*

Problems make use of the detailed data regarding a Phase 2 randomized clinical trial of beta carotene supplementation (bcarot.txt) as documented in the file bcarot.doc. (Make certain you get the correct file, there is also a file carot.txt that is restricted to some summary data from this RCT. Do not use that file.) The file can be read using the following command all on one line (cutting and pasting should work):

```
quietly: infile ptid str10 dbirth age str3 sex str10 race
height weight bsa pctfat tri chol hdl ldl tpro alb str10
dstart str10 drand arm nvisit str10 dvisit tvisit bcarot
ret retpalm vite dose str10 dlast tlast using bcarot.txt
```

(Note that several of the variables are dates that are being read in as character strings. Those variables are preceded with "str10" which tells Stata that the variable is a character string that might be 10 characters long. Similarly, *race* is read in as a character string that might be 10 characters long, and *sex* is read in as a character string that might be 3 characters long.)

You may find it useful to create an indicator variable for females:

```
generate female= .
replace female= 1 if sex=="F"
replace female= 0 if sex=="M"
```

(For this homework, we will not use the dates or race, so we will postpone learning how to convert them to a useful form for data analysis.)

Note that this data file contains two variables related to dose. Variable *arm* tells which group the patient was randomly assigned to, and variable *dose* tells the dose of beta carotene the patient was taking during the weeks just prior to their measurement. (Recall that the first three months, all patients were taking placebo, and after completing the full course of study drug, the patient was no longer taking study drug.

In this homework, we are going to restrict attention to the measurements made while the patient was not actively taking beta carotene supplements. If you wanted, you could just drop all rows that we are not interested in:

```
drop if dose!=0
```

(Some of the requested analyses are not very scientifically based…that is the point.)

Questions for Biost 514 and Biost 517:

1. Using only those measurements made while the patient was not taking beta carotene supplementation (so dose=0), generate appropriate descriptive statistics for the measurements of beta carotene (*bcarot*), retinol (*ret*), retinol palmitate (*retpalm*), and vitamin E (*vite*) by sex.

Ans: Descriptive statistics are contained in Table 1 below.

Table 1: Descriptive statistics for all measurements made while taking no beta carotene supplementation. Statistics provided include the number of available measurements without missing data (N), the number of cases with missing data (msng), the standard deviation (SD), the median (Mdn), the 25th and 75th percentiles (interquartile range or IQR), the minimum (Min), and the maximum (Max).

		N (msng)	Mean (SD)	Mdn (IQR)	(Min, Max)
Males	**Beta carotene (mg/L)**	193 (1)	274 (248)	218 (131, 332)	(34, 1944)
	Retinol (mg/L)	193 (1)	607 (121)	608 (518, 695)	(340, 975)
	Retinol palmitate (mg/L)	193 (1)	27.7 (36.2)	17.0 (10.0, 29.0)	(0, 308)
	Vitamin E (mg/dL)	193 (1)	7.20 (1.90)	7.36 (6.06, 8.35)	(0.00, 12.82)
Females	**Beta carotene (mg/L)**	177 (1)	380 (286)	296 (191, 426)	(44, 1680)
	Retinol (mg/L)	177 (1)	532 (133)	499 (444, 592)	(295, 1042)
	Retinol palmitate (mg/L)	177 (1)	25.2 (21.6)	17.0 (12.0, 31.0)	(0, 100)
	Vitamin E (mg/dL)	177 (1)	7.52 (1.76)	7.87 (6.37, 8.88)	(3.09, 12.27)
TOTAL	**Beta carotene (mg/L)**	370 (2)	325 (272)	252 (156, 386)	(34, 1944)
	Retinol (mg/L)	370 (2)	571 (132)	552 (469, 656)	(295, 1042)
	Retinol palmitate (mg/L)	370 (2)	26.5 (30.1)	17.0 (11.0, 30.0)	(0, 308)
	Vitamin E (mg/dL)	370 (2)	7.35 (1.84)	7.52 (6.16, 8.58)	(0.00, 12.82)

2. Suppose we want to compare the beta carotene levels for men to those for women during the time with no active beta carotene supplementation using only those measurements made while the patient was not taking beta carotene supplementation (so *dose* = 0).

    a. From your results in problem 1, compare the mean beta carotene levels recorded for women in the dataset to the mean of measurements recorded for men. (No statistical test need be performed. Just compare the descriptive statistics.) What scientific question would this be addressing?

Ans: The measurements made on women tend toward slightly higher beta carotene levels as evidenced by their sample mean being 106 mg/L higher than that for men. (This is also evidenced in the median, as well as the quartiles, and the women's measurements appear to be slightly more variable as evidenced by the larger standard deviation and the larger interquartile range- 234 vs 201). However, this comparison answers no important scientific question owing to the variable number of measurements made on each subject and the fact that some measurements were made following several months of beta carotene supplementation at varying doses for some subjects and not for others

    b. Generate appropriate descriptive statistics for measurements of beta carotene levels by sex during the first three months on study (so *nvisit* <= 3). What scientific question would this be addressing?

Ans: Descriptive statistics are contained in Table 2b below. The measurements made on women prior to randomization tend toward slightly higher beta carotene levels as evidenced by their sample mean being 60 mg/L higher than that for men. These statistics are of more interest than those for problem 1, because we have eliminated the aspect of some measurements made following periods of supplementation with beta carotene at varying doses. However, there are still varying numbers of measurements made on each subject, so we are not treating subjects equally.

Table 2b: Descriptive statistics for all measurements made while taking no beta carotene supplementation prior to randomization. Statistics provided include the number of available measurements without missing data (N), the number of cases with missing data (msng), the standard deviation (SD), the median (Mdn), the 25th and 75th percentiles (interquartile range or IQR), the minimum (Min), and the maximum (Max).

		N (msng)	Mean (SD)	Mdn (IQR)	(Min, Max)
Males	**Beta carotene (mg/L)**	87 (1)	197 (125)	169 (105, 254)	(34, 600)
	Retinol (mg/L)	87 (1)	618 (125)	608 (519, 703)	(371, 975)
	Retinol palmitate (mg/L)	87 (1)	32.0 (47.3)	18.0 (11.0, 28.0)	(0, 308)
	Vitamin E (mg/dL)	87 (1)	7.79 (1.65)	8.00 (6.87, 8.66)	(0.00, 12.02)
Females	**Beta carotene (mg/L)**	95 (1)	257 (116)	244 (164, 337)	(44, 715)
	Retinol (mg/L)	95 (1)	531 (133)	488 (444, 583)	(295, 990)
	Retinol palmitate (mg/L)	95 (1)	23.3 (19.1)	17.0 (12.0, 24.0)	(0, 94)
	Vitamin E (mg/dL)	95 (1)	8.16 (1.40)	8.19 (7.38, 9.15)	(5.00, 12.27)
TOTAL	**Beta carotene (mg/L)**	182 (2)	228 (123)	208 (142, 309)	(34, 715)
	Retinol (mg/L)	182 (2)	572 (136)	552 (469, 669)	(295, 990)
	Retinol palmitate (mg/L)	182 (2)	27.5 (35.6)	17.0 (12.0, 26.0)	(0, 308)
	Vitamin E (mg/dL)	182 (2)	7.98 (1.53)	8.14 (7.13, 9.00)	(0.00, 12.27)

c. Compare the mean beta carotene level for women at the start of the study (so *dstart* = *dvisit*) to the corresponding mean for men. What scientific question would this be addressing?

Ans: Descriptive statistics are contained in Table 2c below. Based on the available measurements, the women tend to have a slightly higher beta carotene level than the men, as evidenced by a sample mean 75 mg/L higher than that for the men. Because the subjects are each represented once, these statistics allow us to examine whether there might be systematic trends in the measurements between men and women before taking beta carotene supplementation as a part of the study. Conceivably any such trends might represent differences between the sexes in the way they would absorb, store, metabolize, and/or excrete beta carotene, differences between the sexes in the beta carotene, etc. in their diets, or differences between the sexes in the way that their plasma levels might be associated with a willingness to enroll in a clinical trial, among others. Alternatively, any of the observed differences may just be due to random sampling error from a population in which men and women have similar distributions of the plasma levels. It should be noted that a single man and a single woman are missing measurements at the time of study accrual. If the cause for the missing data is related to the value of the unknown missing measurements, then the data in the table below would be biased for any scientific use.

Table 2c: Descriptive statistics for all measurements made at the time of study accrual. Statistics provided include the number of available measurements without missing data (N), the number of cases with missing data (msng), the standard deviation (SD), the median (Mdn), the 25th and 75th percentiles (interquartile range or IQR), the minimum (Min), and the maximum (Max).

		N (msng)	Mean (SD)	Mdn (IQR)	(Min, Max)
Males	**Beta carotene (mg/L)**	21 (1)	204 (148)	179 (100, 247)	(46, 600)
	Retinol (mg/L)	21 (1)	594 (110)	576 (524, 691)	(403, 779)
	Retinol palmitate (mg/L)	21 (1)	21.6 (16.7)	17.0 (11.0, 23.0)	(10, 74)
	Vitamin E (mg/dL)	21 (1)	8.19 (1.50)	8.39 (7.49, 9.04)	(5.20, 12.02)
Females	**Beta carotene (mg/L)**	23 (1)	279 (122)	257 (164, 381)	(117, 614)
	Retinol (mg/L)	23 (1)	504 (121)	492 (405, 578)	(295, 808)
	Retinol palmitate (mg/L)	23 (1)	24.4 (20.6)	19.0 (11.0, 25.0)	(10, 94)
	Vitamin E (mg/dL)	23 (1)	8.30 (1.59)	8.51 (7.38, 9.27)	(5.00, 11.61)
TOTAL	**Beta carotene (mg/L)**	44 (2)	244 (139)	220 (138, 324)	(46, 614)
	Retinol (mg/L)	44 (2)	547 (123)	526 (451, 669)	(295, 808)
	Retinol palmitate (mg/L)	44 (2)	23.1 (18.7)	18.0 (11.0, 24.0)	(10, 94)
	Vitamin E (mg/dL)	44 (2)	8.25 (1.53)	8.45 (7.44, 9.15)	(5.00, 12.02)

d. How does the difference between the sexes computed in part c compare to the difference you found in parts a and b? Which is more appropriate to address the question of a tendency for women to have different beta carotene levels than men? Why?

Ans: The difference in the sample means for all measurements was 106 mg/L, the difference in the sample means for all measurements made prior to randomization was 60 mg/L, and the difference in the sample means for the single measurement made on each subject at study accrual was 75 mg/L. Because this last comparison is treating each woman and each man as equally important, this is the one that is best addressing differences between men and women in their beta carotene values. Of course, without knowing whether our RCT population represents a random sample among all men and women, we must be cautious in our generalizations.

3. Dealing with longitudinal data is always hard. Even when there are supposed to be measurements made at specific times, in real life it is rare that that will be the case for all patients. Hence, we must take a lot of time to verify the sampling scheme for the study. One of the first tasks is to find out how many distinct patients you have data on.
   a. How many observations are there when the patient is not taking beta carotene supplementation, and how many distinct patients are represented among that group? In Stata this can be effected by looking at the output for number of unique values of *ptid* with commands like

```
codebook ptid
codebook ptid if dose==0
```

Ans: There are 370 available (nonmissing) plasma beta carotene measurements made on 46 unique patients.

4. In problems 1 and 2, you generated descriptive statistics using all measurements made while the patient was not taking beta carotene supplements. However, multiple measurements were made on each subject. This problem guides you through the process of using Stata to determine how many repeat measurements are made on each individual.

   The data file contains repeated measurements on each individual. When our interest is on how individuals fare, we often combine such repeated measurements into a single

summary. For instance, we might consider taking the average of the measurements, the maximum or minimum of the measurements, or only the last measurement. Stata provides a command "egen" that will allow us to easily abstract such summaries by individual.

For instance, suppose we want the mean beta carotene level for each individual. We can obtain a variable *mncarot* that will contain that by:

- `egen mncarot = mean(bcarot), by(ptid)`

Each row will now have a value for variable *mncarot* that is equal to the mean of all the beta carotene values for that individual. If you wanted to have instead the mean of beta carotene levels during the run in period you could use:

- `egen mncarotRI = mean(bcarot) if nvisit <= 3, by(ptid)`

After this command, you would have a variable that had missing values for any rows corresponding to visits after the patient was on run-in, and for all other rows, the value for variable *mncarot* would be equal to the mean of measurements for visits 0 -3 for that individual.

In this and the following problems you will need to use "`egen`" repeatedly in order to be able to perform analyses on a per individual rather than per measurement basis.

a. Use "`egen`" to generate a variable *mncarot* reflecting the mean beta carotene levels for each individual while not taking beta carotene supplements. Provide descriptive statistics by sex and by treatment arm. How do the mean values compare across those groups?

Ans: The descriptive statistics presented in Table 4a suggest that measurements made on women average 106 mg/L higher than those for men.

Table 4a: Descriptive statistics for the mean plasma beta carotene level for each subject computed from all available measurements made when that subject had not been on the investigational beta carotene supplementation during the previous month. Each subject is represented in these statistics in direct correspondence to their number of available measurements. Statistics provided include the number of available measurements without missing data (N), the number of cases with missing data (msng), the standard deviation (SD), the median (Mdn), the 25th and 75th percentiles (interquartile range or IQR), the minimum (Min), and the maximum (Max).

	N (msng)	Mean (SD)	Mdn (IQR)	(Min, Max)
Males	194 (0)	274 (172)	236 (156, 357)	(59, 933)
Females	178 (0)	380 (153)	361 (249, 494)	(155, 800)
Arm 0	133 (0)	258 (105)	204 (156, 357)	(129, 405)
Arm 15	64 (0)	287 (136)	243 (188, 402)	(96, 513)
Arm 30	62 (0)	357 (157)	305 (199, 494)	(179, 617)
Arm 45	50 (0)	422 (214)	499 (228, 516)	(93, 800)
Arm 60	63 (0)	398 (219)	361 (256, 441)	(59, 933)
TOTAL	372 (0)	325 (171)	305 (188, 422)	(59, 933)

b. How do the results in part a of this problem compare to the results you obtained in part a of problem 2?

Ans: The sample means are exactly the same. *(This is due to something called the "double expectation theorem": By computing the mean for each individual (the "conditional expectation" for each individual), and then replacing each original observation with that individual's mean, we will just obtain the overall mean or the original measurements (the "unconditional expectation").)*

The standard deviations in Table 4a are less than those in Table 1. This is because we have removed the within-individual variability. Similarly, the minima and maxima are less extreme in Table 4a owing to the use of individual specific means.

The medians and quartiles are different between the tables in an unpredictable way. *(There is no such thing a "double percentile" formula: It is in general impossible to combine medians and quartiles from different samples in a way that would allow us to estimate the median and quartiles in the total sample.)*

c. Use "egen" to generate a variable *visits* counting the number of visits with available data for each individual while not taking supplements, and provide suitable descriptive statistics for this variable using all cases while not using supplements, as well as by sex and by treatment arm when not using supplements. How do these analyses affect the scientific use you would make of the analyses in part a of this problem and/or part a of problem 2? The following Stata code will generate a variable *visits*:

```
egen visits= count(bcarot), by(ptid)
```

Ans: The descriptive statistics presented in Table 4c suggest that women average fewer measurements than do men, and there are markedly more measurements for each individual in the placebo group than for the other arms. This suggests that we need to be cautious when comparing descriptive statistics: statistics that are heavily dependent on sample sizes (such as minimum and maximum) may be uninterpretable. Furthermore, the standard deviations can be misleading. It should also be noted that the numbers in Table 4c are themselves biased as they over represent the subjects having more measurements.

Table 4c: Descriptive statistics for the number of visits for each subject computed from all available measurements made when that subject had not been on the investigational beta carotene supplementation during the previous month. Each subject is represented in these statistics in direct correspondence to their number of available measurements. Statistics provided include the number of available measurements without missing data (N), the number of cases with missing data (msng), the standard deviation (SD), the median (Mdn), the 25th and 75th percentiles (interquartile range or IQR), the minimum (Min), and the maximum (Max).

	N (msng)	Mean (SD)	Mdn (IQR)	(Min, Max)
Males	194 (0)	10.8 (4.9)	7 (7, 16)	(4, 18)
Females	178 (0)	8.7 (4.0)	7 (7, 7)	(4, 18)
Arm 0	133 (0)	15.5 (2.7)	16 (15, 18)	(7, 18)
Arm 15	64 (0)	6.5 (1.0)	7 (7, 7)	(4, 7)
Arm 30	62 (0)	6.9 (0.3)	7 (7, 7)	(6, 7)
Arm 45	50 (0)	6.4 (1.1)	7 (6, 7)	(4, 7)
Arm 60	63 (0)	6.5 (1.0)	7 (7, 7)	(4, 7)
TOTAL	372 (0)	9.8 (4.6)	7 (7, 15)	(4, 18)

d. Doing descriptive statistics on the summarized variable is still complicated due to the number of repeated measurements on each individual. If we want to find out the distribution of *mncarot* across individuals (rather than rows in the file), we will need to restrict our analysis to one row for each patient. In this data set every subject should have had a measurement on their start date (I had you use that idea above, but we did not verify that it was valid). How many subjects have a measurement in which *dvisit = dstart?* (You could use the `codebook` command again.)

Ans: Each of the 46 unique subjects have one visit corresponding to the starting date.

e. Now, since we know that every individual has a row corresponding to *dvisit = dstart*, when we desire statistics on each individual, we could obtain summary statistics just for rows corresponding to *dvisit == dstart*. Describe the distribution of the mean beta carotene level for each subject while not taking beta carotene supplements, making sure to represent each subject once. Also provide statistics within strata defined by sex and within strata defined by assigned treatment arm. What scientific use would you make of these analyses?

Ans: The descriptive statistics presented in Table 4e suggest that the mean measurements made on women average 102 mg/L higher than those for men. There is also evidence of a trend toward higher average values (and median, quartiles) for subjects in the arms randomized to higher doses. This is likely due to the fact that we include measurements made relatively soon after supplementation has been stopped. To the extent that it might take some days or weeks for plasma levels to return to normal after stopping supplementation, these statistics are influenced by any effect of supplementation on plasma levels. Because of this, there is no important scientific question answered by this analysis.

Table 4e: Descriptive statistics for the mean plasma beta carotene level for each subject computed from all available measurements made when that subject had not been on the investigational beta carotene supplementation during the previous month. Each subject is represented in these statistics only once. Statistics provided include the number of available measurements without missing data (N), the number of cases with missing data (msng), the standard deviation (SD), the median (Mdn), the 25th and 75th percentiles (interquartile range or IQR), the minimum (Min), and the maximum (Max).

	N (msng)	Mean (SD)	Mdn (IQR)	(Min, Max)
Males	22 (0)	280 (192)	238 (169, 357)	(59, 933)
Females	24 (0)	382 (159)	381 (245, 497)	(155, 800)
Arm 0	9 (0)	256 (112)	204 (156, 357)	(129, 405)
Arm 15	10 (0)	288 (141)	243 (188, 402)	(96, 513)
Arm 30	9 (0)	354 (165)	305 (199, 494)	(179, 617)
Arm 45	8 (0)	394 (232)	418 (198, 512)	(93, 800)
Arm 60	10 (0)	381 (229)	346 (256, 441)	(59, 933)
TOTAL	46 (0)	333 (181)	289 (192, 460)	(59, 933)

5. In problem 4, you generated descriptive statistics using all measurements made while the patient was not taking beta carotene supplements. This problem considers measurements made while on run-in.
   a. Use "`egen`" to generate a variable *nrunin* counting the number of visits with available data for each individual while on run-in, and provide suitable

descriptive statistics for this variable using all cases in the datafile. The following Stata code will generate a variable *grbg*:

```
egen grbg= count(bcarot) if nvisit<=3, by(ptid)
```

This variable will have missing data for any case in which *nvisit* was not less than or equal to 3 (i.e., not during run-in). You can try looking at the data for the first patient to see this:

```
list bcarot grbg if ptid==701
```

In order to obtain an entry for every row in the data set, we again use "egen":

```
egen nrunin= mean(grbg), by(ptid)
```

Again, try looking at the case to see what happened:

```
list bcarot grbg nrunin if ptid==701
```

Even after this command, there would still be missing data for any individual who never had a run-in visit. (This is not a problem in this dataset, but it is a good idea to think about it in all cases.) Hence, we can now enter 0 for all cases in which *nrunin* is missing.

```
replace nrunin= 0 if nrunin==.
```

How many measurements in the datafile correspond to a patient having fewer than four run-in measurements? How many individuals does this represent? How many individuals have more than four run-in measurements? You might consider the following Stata commands:

```
table nrunin
codebook ptid if nrunin<4
```

Ans: Twenty (20) rows in the datafile correspond to subjects with only three values during the run-in phase, but this represents only three subjects. One subject has more than four run-in measurements.

b. We are interested in how beta carotene levels might vary across individuals prior to treatment. Describe the distribution of mean beta carotene levels for each subject during the run-in period. Provide statistics overall and within strata defined by sex and within strata defined by arm. Comment on any problems there might be in the interpretation of these descriptive statistics.

Ans: The descriptive statistics presented in Table 5b suggest that the mean measurements made on women average 62 mg/L higher than those for men. There is also a slightly higher sample mean for patients on the placebo arm relative to those randomized to the other arms (with quite similar values for the run-in plasma levels for all of the other arms). Because all subjects are similar with regard to not receiving supplementation and because each subject is represented once, these statistics are generally useful to describe the average plasma beta carotene levels for the subjects as grouped by sex or race. It should be noted that these statistics pertain to the average of multiple measurements made on each subject, and thus the standard deviations and the minima and maxima reflect summarization of averages rather than summarization of a single measurement. Furthermore, the slight variation in the number of measurements per subject will cause the standard deviations and the extrema to be less comparable.

Table 5b: Descriptive statistics for the mean plasma beta carotene level for each subject computed from all available measurements made during the run-in period prior to randomization. Each subject is represented in these statistics only once. Statistics provided include the number of available measurements without missing data (N), the number of cases with missing data (msng), the standard deviation (SD), the median (Mdn), the 25th and 75th percentiles (interquartile range or IQR), the minimum (Min), and the maximum (Max).

	N (msng)	Mean (SD)	Mdn (IQR)	(Min, Max)
Males	22 (0)	198 (117)	175 (135, 241)	(48, 476)
Females	24 (0)	259 (103)	233 (177, 347)	(98, 496)
Arm 0	9 (0)	260 (131)	223 (149, 360)	(136, 476)
Arm 15	10 (0)	222 (131)	186 (136, 238)	(65, 496)
Arm 30	9 (0)	224 (88)	233 (140, 282)	(126, 349)
Arm 45	8 (0)	227 (106)	216 (148, 299)	(93, 396)
Arm 60	10 (0)	218 (122)	224 (98, 311)	(48, 408)
TOTAL	46 (0)	230 (113)	212 (140, 311)	(48, 496)

c. How do the standard deviations observed in part b of this problem compare to analogous results observed in part c of problem 2?

Ans: The standard deviations of 117 and 103 in Table 5b are smaller than those of 148 and 122 in Table 2c. We must make a distinction between the variation we might see in a population of measurements taken one per person and the variation we might see when we average several measurements for each individual. In essence, there is variability <u>within</u> each individual around his/her personal average. This variability is attenuated when we take the average of several measurements. But there is also variability <u>between</u> individuals in the personal averages. When we look at the SD of the averaged measurements in Table 5b, we are getting closer to an estimate of the between individual variance. In Table 2c, we are measuring the combination of the within and between individual variance.

6. Repeat the types of analyses you performed in problem 5b, but now consider the maximum beta carotene measurement for each subject. What problems does the sampling scheme present for this problem. A variable *mxcarot* can be created using

```
egen mxcarotRI = max(bcarot) if nvisit <= 3, by(ptid)
```

Ans: The descriptive statistics presented in Table 6 suggest that the maximum measurement made on each woman averages 70 mg/L higher than those maxima for men. There is little evidence of a trend toward higher average values (and median, quartiles) for subjects in the across randomization arms. It should be noted that the variation in number of measurements taken for each individual may cause noncomparability of the maxima. However, to the extent that the number of measurements taken is independent of the measurements themselves <u>and</u> the groups being compared, this may not be such a huge problem if our only interest is looking at how comparable the groups are. *(In general, however, we always get nervous looking at extreme values when we might have more measurements for some subjects than other. A common setting that causes particular problems is when repeat measurements are made precisely because large or small values were seen.)*

Table 6: Descriptive statistics for the maximum plasma beta carotene level for each subject computed from all available run-in measurements made when that subject had not been on the investigational beta carotene prior to randomization. Each subject is represented in these statistics only once. Statistics provided include the number of available measurements without missing data (N), the number of cases with missing data (msng), the standard deviation (SD), the median (Mdn), the 25th and 75th percentiles (interquartile range or IQR), the minimum (Min), and the maximum (Max).

	N (msng)	Mean (SD)	Mdn (IQR)	(Min, Max)
Males	22 (0)	252 (135)	240 (169, 273)	(65, 600)
Females	24 (0)	322 (144)	301 (219, 401)	(117, 715)

Arm 0	9 (0)	305 (156)	258 (190, 385)	(160, 600)
Arm 15	10 (0)	300 (203)	240 (169, 279)	(87, 715)
Arm 30	9 (0)	278 (100)	255 (199, 326)	(154, 428)
Arm 45	8 (0)	290 (107)	263 (231, 386)	(123, 437)
Arm 60	10 (0)	272 (144)	291 (117, 388)	(65, 488)
TOTAL	46 (0)	289 (143)	258 (190, 385)	(65, 715)

7. Repeat the types of analyses you performed in problem 5, but now restrict attention to beta carotene measurements made for each individual after stopping study drug (so *nvisit* > 12 and verify that *dlast* is missing for all these measurements). What scientific question is this addressing? Is there evidence of outliers in the data? Explain.

Ans: The descriptive statistics are presented in Table 7 suggest for the average measurement made on each subject following discontinuation of beta carotene supplementation. It should be noted that the higher rate of missing data for women over men raises fears that the generalization of these results may be biased. There is no striking trend in missingness by treatment arm, however that does not preclude different mechanisms of missingness across treatment arms that could lead to bias.

Among the measurements that we do have, there appears to be a trend toward higher individual averages among those subjects who were randomized to higher doses. This is suggestive of a carry-over effect of the supplementation that might last a month or more, but more detailed analysis would be needed to better describe time trends of any such carry-over. Interpretation of the differences between sexes may be confounded by the treatment assignment: At randomization, a slightly lower proportion of women were assigned to placebo, and the patterns of missing data may have accentuated any effect or attenuated the effect.

Table 7: Descriptive statistics for the mean plasma beta carotene level for each subject computed from all available measurements made after that subject had stopped investigational beta carotene. Each subject is represented in these statistics only once. Statistics provided include the number of available measurements without missing data (N), the number of cases with missing data (msng), the standard deviation (SD), the median (Mdn), the 25th and 75th percentiles (interquartile range or IQR), the minimum (Min), and the maximum (Max).

	N (msng)	Mean (SD)	Mdn (IQR)	(Min, Max)
Males	20 (2)	416 (357)	290 (238, 528)	(110, 1634)
Females	19 (5)	607 (298)	565 (402, 838)	(89, 1340)
Arm 0	8 (1)	224 (139)	185 (119, 306)	(89, 486)
Arm 15	8 (2)	396 (240)	315 (252, 494)	(137, 908)
Arm 30	9 (0)	533 (277)	402 (332, 792)	(231, 974)
Arm 45	6 (2)	760 (378)	785 (491, 914)	(246, 1340)
Arm 60	8 (2)	692 (397)	584 (507, 677)	(363, 1634)
TOTAL	39 (7)	509 (339)	450 (258, 696)	(89, 1634)