SURVEY SAMPLING: ANSWERS & DISCUSSION

Once again, the A&D answers are considerably more detailed and discursive than those you were expected to provide. This is typical of the *Answers and Discussion* that will be attached to your returned Problems Sets. Remember that you should always read these attachments with care, even if you got a top grade on the Problem Set. Occasionally the A&Ds will introduce supplementary course material. You should regard the Answers and Discussion attachments as basic course material — similar to the Problem Sets themselves, the Course Pack handouts, PowerPoints, and assigned readings.

I.   (a)   **1,000/17,250 ≈.058**

(b)   **1,000/17,250** ≈ .058 (= 5.8%) [in any *simple random sample*, the answers to (a) and (b) are the same]

(c)   847/1000 = **.847** (or **84.7%%**)

(d)   407/847 = **48.1%**

(d)   Sampling error is based in the size of the *completed* sample (i.e., $n = 847$), not the drawn sample of $n = 1,000$ (for which the margin of error would be about $\pm 3.2\%$). Refer to the table on p. 72 of Weisberg et al., in which $n = 750$ is the sample size shown that is closest to the completed sample size in the present problem. The sampling error for a simple random sample of $n = 750$ is given as 3.6%%, so sampling error for n = 838 would be a bit smaller, maybe 3.5% or 3.4%. Alternatively use the approximate formula given in the handout:

$$\text{margin of error (95\% confidence interval)} \approx \pm 100\% / \sqrt{n}$$

Since $\sqrt{838} \approx 29.1$, sampling error $\approx \pm 100\% / 29.1 = \pm 3.44\%$, so the table and formula agree (as they should).

Note. Since the sampling fraction is fairly large and sampling is (presumably) without replacement, sampling error is actually somewhat less than that shown in the table or given by the formula.

(e)   **Probably Not.** The *sample statistic* is **48.1%** of students in the completed sample approve of President's Bush's performance as President (while $440/847 \approx 51.9\%$ disapprove), so it is true that there are somewhat more disapprovers than approvers in the sample. But, as an estimate of the (unknown) *population parameter*, i.e., the proportion of disapprovers in the whole student population, this statistic is subject to a sampling error of about $\pm 3.44\%$, which is greater than the 1.9% difference between 48.1% and a bare majority of 50%. This means that we can be 95% confident only that the true population parameter (percent of approvers) lies between about 44.7% and 51.5% — an interval that includes parameter values that entail more approvers than disapprovers. [However, if the population parameter actually lies outside of the 95% confidence interval of about 45%–52%, it is equally likely to be lower than 44.7% as to

be higher than 51.5%. It turns out (if we consult a more detailed statistical table) that we can be about 90% confident that at least 50% of all students disapprove of the President's performance.] There is also a further problem: the response rate fell well below 100% and it is possible that we disproportionately failed to interview approvers.

II   APPROVE OF PRESIDENT'S JOB PERFORMANCE is Question 12 in the Student Survey, so you must count up the coded values in column 9 of the data spreadsheet you were provided. You should come up with following (which we will soon learn to call a *frequency distribution*):

Code	Frequency	Label
1	4	Strongly Approve
2	22	Approve
3	15	Disapprove
4	3	Strongly Disapprove
5	6	No Opinion
9	0	
Total	50	

So the population parameter (percent of the population that approves) is 26/50 = **52%**. In each sample of 10, count up the number of respondents coded either 1 of 2 and divide by 10 to get the sample statistic. Obviously different students drew different random samples that produced different sample statistics.

Alternatively, we might define the population parameter as the percent of approvers among everyone who either approves or disapproves (excluding people with no opinion). The population parameter is then 26/44 ≈ **59%**. In each sample of 10, count up the number of respondents coded either 1 of 2 and divide by the number of respondents in the sample (perhaps less than 10) who either approve or disapprove to get the sample statistic, excluding those with no opinion.

Sampling with and without replacement. This raises the question of what to do if (for example) a Table of Random Numbers produces the same number (corresponding to an actual case) twice (or more) during the drawing of a single sample. One option is to discard the number on the second (and any subsequent) draw, in which case you are *sampling without replacement*. The other option is to allow the same case to be drawn into a sample more than once (and to count it accordingly in sample statistics), in which case you are *sampling with replacement*. Sampling without replacement is more common and produces (at least slightly) smaller sampling error for any sample size. On the other hand, formulas and tables for sampling error typically assume sampling with replacement, because the underlying calculations are then much simpler. (See the bottom of the Theoretical Probabilities handout.) In any case, if the sampling fraction is less than about 1/100 (and it is *much* less than this in almost all survey research), there is no practical difference between the two modes of sampling. Here, of course, your sampling fraction is a quite large (10/50 ≈ 20%). A bit of experimentation with the Simple Random Sample applet should make it clear that it samples *without replacement*. (Set

the population size at 10 and select repeated samples of size 5; you always find five different numbers in the sample bin and the other five remaining in the population hopper.)

Note. Bear in mind that what is being said in this respect is that no case will appear more than once in any *single* sample of size $n = 10$. Of course, a given case may appear in several *different* samples of size $n = 10$ — indeed, if you take 10 samples of size $n = 10$ out of a population of size $N = 50$, there must be a lot of duplications of the latter type.

By the approximate formula, the margin or error for a sample of size $n = 10$ is about $100\%/\sqrt{10} \approx$ **32%**. Remember that these calculations assume sampling with replacement, so if you sampled with replacement (as you probably did), your samples are subject to somewhat smaller (but still very large) sampling error.

You will probably (but not necessarily) find that the spread between your largest and smallest sample statistics from ten samples of size $n = 10$ each to be on the order of 30-50 percentage points. If you took many a great many samples of size $n = 10$, you would find their samples statistics would average out to just about the population parameter of about 52% (or 59%), even though individual samples can give only the statistics $0/10 = 0\%$, $1/10 = 10\%$, $2/10 = 20\%$, $3/10 = 30\%$, etc. Moreover, you would find that about 95% of them would be between about 20% and 85% — that is (approximately) $52\% \pm 32\%$. This is not very informative, of course, which is why we almost always use samples considerably larger than $n = 10$. Of course, you can pool your 10 samples of 10 into a single pooled sample of $n = 100$; the sample statistic based on the pooled sample has a margin of error of about $100\%/\sqrt{100} \approx$ **10%**. Note that whether or not the individual samples were taken with replacement, the pooled sample necessarily entails replacement (otherwise the pooled sample size of $n = 100$ could not exceed the population size of $N = 50$).

III.

1   You shouldn't be convinced that a majority of the members constituents oppose the bill. The *population* of interest is all (adult) constituents in the member's district. The (unknown) population *parameter* of interest is the percent or proportion of constituents who oppose the bill that would provide government sponsored insurance for nursing home care. The *sample* consists of the 1128 letter-writers among the constituents, and the (known) sample *statistic* is the $871/1128 \approx 77\%$ of the sample of letter-writing constituents who oppose the bill. But the sample is entirely *self-selected*, and we may reasonably conjecture that only people who have quite strong opinions on an issue will take the trouble to write to their representatives about it. It may be that more conservative "anti-government" ("Tea Party") constituents, who are more likely to be in opposition, are more likely to express their views (and to be aware of the proposal in the first place and motivated to write a letter about it) than many of its "natural" supporters among the poor, elderly, or infirm. Moreover, business-oriented interest groups may have mobilized their members to write letters opposing the bill. (Of course, others, e.g., nursing homes that would in effect be subsidized by the proposed insurance, might also mobilize the other side.) Note that, even if the population parameter could be accurately estimated using a representative sample, it may actually be expedient for representatives to follow "letter opinion" (the known but probably biased sample statistic) rather than overall district opinion, since the considerations that suggest "letter opinion" is biased also suggest that the letter writers are more likely than non-letter writers to vote on the basis of this issue in future elections.

2.   (a)   The population referred to in the newspaper report is all readers of the newspaper, though the paper may believe that this is about the same as all residents of the community.

     (b)   A lot of students said that the true proportion of readers/residents who favor one-way streets is likely to be larger than 14/98 because "complainers" dissatisfied with a recent policy change are more likely take the trouble to call. This may well be true. Note that this argument implies that, if the one-way streets were later converted back to two-way, the balance of opinions expressed in call-in polls would reverse. However, there may be an argument that suggests that call-in opinion may consistently underestimate support for one-way streets. One-way streets expedite the general flow of traffic (and may allow higher speed limits). Converting some two-way streets into one-way streets thus probably benefits most drivers (and residents) of the town a bit. But each of the many beneficiaries is benefitted only slightly and so are unlikely to make the effort to respond to a voluntary newspaper survey. However, residents who live along the streets in question likely feel directly, substantially, and negatively affected by the change: they may have to take a more circuitous routes to leave or return home and they almost certainly experience more, faster moving, and noisier traffic in front of their houses. Plausibly these residents are quite strongly opposed, and therefore they are quite likely to respond to a voluntary-response newspaper survey. (This is the typical NIMBY — "not in my back-yard" or, in this case, "not along my front-yard" — phenomenon, and the intense opinions in opposition are likely not only to dominate a voluntary-response survey but also to win out politically.)

3.   Respondents to any "call-in" survey are self-selected. Very predictably, self-selected respondents have more intense opinions on the subject of the survey than those who do not chose to respond. (Indeed, there may be nothing to stop those with intense views — or those who can be mobilized by advocacy organizations — from "stuffing the ballot box" by calling in many times.) Probably the relatively small number of committed UN critics are more intense in their views than the much greater numbers of people who are at least marginally favorable to the UN.   A nationwide random sample of 500 respondents has a sampling error of about $100\% / \sqrt{500} \approx \pm 5\%$, so we can be 95% confident that at least 68% of the general population would answer "yes." (Actually, we can be 97.5% confident of this — can you see why? [Look back at the discussion of I(e).]) On the other hand, the self-selected sample, despite its much greater size, is likely to be highly biased, for the reasons suggested above.

4.   (a)   Probability that a male faculty member is selected = 100/1000 = **0.1**
           Probability that a female faculty member is selected = 50/500 = **0.1**

     (b)   Given a *simple random sampling* procedure, every sample of 150 faculty members is equally likely to be selected, including samples that *do not* include exactly 100 men and 50 women. But under the sampling procedure described, only samples that include *exactly* 100 men and 50 women have a chance to be selected. The result is a random sample that is *stratified* (by gender), which (with respect to parameters on which there are male-female differences) has a slightly smaller sampling error than an SRS of the same size. (*Note*: a *stratified* random sample is different from either a *systematic* random sample or a *multistage* random sample.)

Note. If the objective of the survey is to compare the responses of male and female faculty members, it would make sense to select a random sample that is stratified by gender *and that has an equal number of men and women*, i.e., 75 of each (if we want to maintain the overall sample size of $n = 150$).   If we did this, sample statistics from the male and female subsamples would have the same margin of error. Of course, if we do this, male respondents must then be *weighted* twice as heavily as female respondents (compensating for the fact that the female sampling fraction is twice the male one) to produce unbiased *overall* sample statistics.

5.   The *population* is all Minneapolis households. The population *parameter* of interest is the percent of households that bake their own bread. The *drawn* sample of 500 may be representative of the population. But the *completed* sample exhibits *availability bias* by including only those households in which someone was home during normal working hours — typically those with a "non-working" (outside the house) wife or older retired persons, who are more likely than others to have the time and inclination to bake. Thus the direction of the bias will almost certainly be to overestimate the proportion of households who bake. (Some students seemed to say that the percent who bake would be underestimated, because people who bake but who also are out of the house working during weekdays will not be found and counted. This is true of course, but people who don't bake but who also are out of the house working during weekdays also will not be found and counted. And for the reasons noted above, the proportion of people who bake is likely to be greater among the stay-at-homes who are counted than among those who work outside of the home and who are not counted.)

6.   The *population* of interest is African-American residents of Miami.   The population *parameter* of interest is (something like) the proportion of the population that is (dis)satisfied with police service. However, the (random) *sample* is drawn only from adults "in predominantly black neighborhoods" so, if the opinions of *all* African-American residents is truly what is to be estimated, minors as well as adults should be interviewed.   More significantly, some (adult and minor) residents of predominantly black neighborhoods are non-black (though such potential respondents presumably could be screened out by interviewers). Much more importantly, *all African-Americans who live in racially mixed neighborhoods are entirely excluded from the sample*, i.e., the sampling frame does not match the population of interest. African-American residents in predominantly black neighbor-hoods may have rather different attitudes about the police from African-American residents in racially mixed neighborhoods (possibly because police behavior varies according to the nature of the neighborhood). In Miami (and many U.S. cities), I suppose we might expect black residents of pre-dominantly black neighborhoods would have somewhat more negative views of the police than black residents of racially mixed neighborhood. (Also, as many of you noted, using uniformed police officers to interview respondents might push responses to be somewhat more favorable to the police — though using black officers might counteract this. But note that this is an *interviewing*, not *sampling*, problem, i.e., it would not be solved by taking a census rather than a sample. *Note*: ANES and similar surveys use interviewers who are (i) women, (ii) professionally dressed but not in any kind of uniform, and (iii) usually of the same race as the respondent.

7. | | *Sample statistic* | *Population parameter* |
|---|---|---|
| (a) | 4.5% | --- |
| (b) | 2.515cm | 2.503cm |
| (c) | 43 | 52% |
| (d) | 73% | 68% |

8. (a)   Sample proportion (statistic) = 702/1190 ≈ **59%**.

   (b)   The (unknown) population parameter of interest is the proportion of the population (U.S. VAP or whatever) that prefers balancing the budget over cutting taxes.

   (c)   We can be 95% confident that the population parameter lies somewhere between 55% and 63% and extremely (but not 100%) confident that it lies above 50%. (Evidently this is not a simple random sample, since an SRS of this size would have a margin of error of about ±3%.)

   Note. In this and most other problems, the population parameter is a *percent* (or proportion), not a *count*. Sample data by itself can provide *no estimate of population counts*, though of course such counts can be estimated by using the sample statistic in conjunction with other information on the size of the population. For example, it is regularly reported that approximately 50 million Americans lack health insurance. Presumably this count is based on a surveys (perhaps the Current Population Survey) in which about 16% of respondents report that they lacked health insurance. Multiplying this sample statistic with the (approximately) known U.S. population of about 310 million produces the 50 million count. Remember that the sample statistic is subject to sampling error. If the statistic comes from the CPS with n = 50,000, the margin of error is about ±0.5%, so we can be 95% confident that the percent of the population without health insurance is between about 15.5% and 16.5% (or, as a count, between 48.1 million and 51.2 million).

9. This questions tests whether students have understood the "counter-intuitive" implication (see slides #30-31 on Random Sampling) that, for most practical purposes, the magnitude of sampling error depends on absolute sample size, and not on the sampling fraction. Evidently many students no not yet understand this.

   (a)   **No** (for all practical purposes). *Sampling error depends on absolute sample size* (as long as the sample is small compared with the population, as is true here even for the smallest states such as Wyoming), *not on the sampling fraction*. Absolute sample size here is a constant n = 2000 over all states, giving a constant margin or error of about ± 2.2%

   (b)   **Yes.** If the sampling fraction/proportion is a constant 1/1000 over all states, absolute sample size will vary from about 525 in Wyoming to 33,000 in California and sampling error will also vary (inversely with the square root of sample size). (The California sample, 48 times larger than the Wyoming sample, therefore would have a sampling error of about $1/\sqrt{48} ≈ 1/7 ≈ 0.14$ the size of the sampling error of the Wyoming sample.)

10.   Chance of a given person appearing in a single poll:

$$1500/150{,}000{,}000 = 1/100{,}000 = \textbf{0.00001} \text{ (or } \textbf{0.001\%}\text{)}$$

Chance of a given person appearing in any one of 20 such polls is, for all practical purposes, 20 times as great or:

$$20/100{,}000 = 1/5{,}000 = \textbf{0.0002} \text{ (or } \textbf{0.02\%}\text{)}$$

Note: The above calculation assumes (incorrectly) that no one can appear in more than one of the 20 polls. For mathematical purists (which probably doesn't include many POLI majors):

chance of *not* appearing *in any one poll* = 0.99999
chance of *not* appearing *in any of 20 polls* =
$$(0.99999)^{20} = 0.999800019$$
chance of appearing in *at least one poll out of 20*
$$= 1 - 0.999800019 = \textbf{0.000199981}$$

The pollsters' samples almost certainly included about the right proportion of Wallace supporters. But the calculations above indicate that in any Wallace campaign rally crowd of (say) about 500-10,000 people, we could expect to find at most only a handful of people who had been (or would be) a respondent in any of the 20 polls. So invariably about 99.98% of any crowd could quite truthfully shout back "No" or "Never" in response to Wallace's question. (Of course, the same would be true in any Nixon or Humphrey 1968 campaign rally crowd.)

Essentially the same would have been true in any McCain or Obama campaign rally crowd in 2008; however, there are now a lot more than 20 polls per election, so the chance of any person appearing in at least one of their samples is more like 1/1,000.)

However, even if all the Gallup and other samples included about the right of Wallace supporters, some of them may have been reluctant to reveal such a voting intention to interviewers, since Wallace was widely depicted as an extreme and "unrespectable" candidate in the national media. Thus there may have been considerable *non-sampling error* in measuring Wallace support. More specifically, among white Southerners (where support for Wallace was quite "normal"), true Wallace supports were probably quite willing to disclose themselves as such. But among white non-Southerners, and especially among more middle-class and better educated ones (where support for Wallace was highly "deviant"), quite a few of the (small number of) true Wallace supporters may have been unwilling to disclose themselves as such until they reached the privacy of the polling place.