Economics 300-01: Quantitative Methods in Economics
Wesleyan University, Fall 2007

Selected Answers to Problem Set #1

1-6: 1. The placebo is to prevent the subjects of the study from knowing whether they are in the control group or the treatment group, thereby precluding any possibility of their behavior changing because of the group to which they were assigned. It also may preclude the administrators of the experiment (e.g. nurses) from knowing who is treated and who is not, so as to prevent any bias in their reporting of the effects.

2. The proportion of people in our sample who received vitamin C treatment and reported being free of colds is $P = 0.26$. The sample size is 400. To calculate the 95% confidence interval for the proportion of people in the *population* who would be free of colds based on this sample, we first find the standard error for the sample proportion estimate, P:

$$\begin{aligned} SE_P &= 1.96\sqrt{\frac{P(1-P)}{n}} \\ &= 1.96\sqrt{\frac{(0.26)(0.74)}{400}} \\ &= 0.043 \end{aligned}$$

Thus we would conclude that, with 95% certainty, between 22% and 30% of the population would benefit from taking vitamin C.

3. We can undertake the same calculation for the sample treated with the placebo. In this case, $P = 0.18$ and $n = 400$. Thus,

$$SE_P = 1.96\sqrt{\frac{(0.18)(0.82)}{400}} = 0.038$$

The standard error is also 4%, making this confidence interval from 14% to 22%. In other words, in the population there is a 95% chance that between 14 and 22% of individuals will never get a winter cold despite not taking vitamin C supplements.

4. This study provides some evidence that taking vitamin C will reduce one's chance of being cold free all winter, but it is not overwhelming because of sampling uncertainty. At one extreme, it is possible that 14% of people will remain cold-free all winter without vitamin C, whereas 30% of people taking vitamin C will avoid a winter cold. At the other extreme, it is possible that one's chance of avoiding a winter cold is 22%, whether or not one takes vitamin C. (We'll further discuss evaluating these confidence intervals in later chapters.)

1-14: 1. Unless the identical car were available with and without seat belts, we would prefer to only look at data from cars with seat belts. That is because cars without seat belts might have other factors that make them more (or less) likely to cause injury or death in an accident. It might be difficult to account for all of these confounding factors (although, as we will see later in the course, it may be possible with multiple regression) so we would prefer to leave those cars without seat belts out of our study.

(Of course, if we could get data from the same make, model and year of car in which some people were wearing seat belts while others were not, that would be an ideal sample — it would allow us to control for all the confounding factors related to the car. But there are other confounding factors that also might matter, such as characteristics about the drivers of cars with and without seat belts. Can you think of an explanation along these lines?)

2. If we could distinguish whether other occupants were wearing seat belts or not, were in the front or back seat, and so on, then it might be useful to pool the data on all occupants — since they were all in the same car driven by the same driver, on the same road, at the same date and time, etc., we could eliminate a number of confounding factors. (This approach still might require regression analysis.) However, it would probably be better to report different statistics for different passenger locations in the car, including the driver, since different locations are more or less likely to result in certain injuries or death. Hence, having data only on drivers would minimize this confounding factor.

3. It would make more sense to separate injuries by level of severity, as seat belts might reduce the severity of certain injuries (say head trauma), and we would miss that information if all injuries were just lumped together.

4. At face value, we would prefer the doctor not to be informed as to whether the patient was wearing a seat belt or not, since that will prevent any bias in the doctor's subjective assessment of the patient's injuries. However, it is difficult to believe that this information would actually change the way a doctor evaluated injuries, so in this case it probably does not matter.

2-14: 1. Notice: $\overline{X} = \frac{1}{n}\sum X_i \implies n \cdot \overline{X} = \sum X_i$.

For case (i), $n_1 = 30$ and $\overline{X}_1 = 100$, thus $\sum_{i=1}^{n_1} X_{1i} = 3000$. Similarly, $n_2 = 70$ and $\overline{X}_2 = 110$, thus $\sum_{i=1}^{n_2} X_{2i} = 7700$. Pooling the two samples, $n = n_1 + n_2 = 100$, $\sum_{i=1}^{n} X_i = \sum_{i=1}^{n_1} X_{1i} + \sum_{i=1}^{n_2} X_{2i} = 10700$, so $\overline{X} = \frac{1}{n}\sum X_i = \frac{10700}{100} = 107$.

Following the same logic, we find:

(ii) $\overline{X} = \frac{10200}{100} = 102$.

(iii) $\overline{X} = \frac{10500}{100} = 105$.

(iv) $\overline{X} = \frac{3150}{30} = 105$.

2. True, and here's a short sketch of the proof:

$$\begin{aligned}
\overline{X} &= \tfrac{1}{n}\sum_{i=1}^{n} X_i = \left(\tfrac{1}{n_1+n_2}\right)\sum_{i=1}^{n} X_i \\
&= \left(\tfrac{1}{n_1+n_2}\right)\left(\sum_{i=1}^{n_1} X_{1i} + \sum_{i=1}^{n_2} X_{2i}\right) \\
&= \left(\tfrac{1}{n_1+n_2}\right)\left(n_1 \cdot \overline{X}_1 + n_2 \cdot \overline{X}_2\right) \\
&= \left(\tfrac{n_1}{n_1+n_2}\right)\overline{X}_1 + \left(\tfrac{n_2}{n_1+n_2}\right)\overline{X}_2
\end{aligned}$$

Notice that the mean of the pooled sample is the weighted mean of the individual sample means, with the weights equal to the relative size of each sub-sample.

2-36: *Historical context*: the U.S. inflation rate was fairly high from the mid-1970s through 1981 — on the order of 10% per year or more — due to a variety of adverse conditions in the economy, including very high oil prices. Starting in late 1979, the Federal Reserve, under the chairmanship of Paul Volcker, adopted a generally contractionary policy stance in an attempt to bring the inflation rate under control. By the end of 1982 they had some initial success, but there was a lot of uncertainty among economists and business people as to whether this decline was permanent or not. (As we now know, it largely was permanent.)

We are asked to find the average forecast for inflation, but are told only ranges of forecasts. Since we do not have enough information to know the actual average for each range, we have to make our best guess about the average for each group of forecasters. The most defensible choice is probably the midpoint of each range, which is what we use below.

1. We can calculate the average inflation forecast with the formulas for grouped data. (See section 2-6 in the text.) For the first group of forecasters, the average forecast (i.e. midpoint of the range) is 3%. 12% of the forecasters held that view. Similarly, 60% forecasted an average inflation rate of 5%, 23% forecasted an average inflation rate of 7%, and 5% forecasted an average inflation rate of 9%. To find the overall average forecast, we take the weighted average of these numbers, with the weights determined by the proportion of forecasters in each group. (Notice that $12\% + 60\% + 23\% + 5\% = 100\%$.)

   Letting $\overline{X}$ be the overall average, we can calculate $\overline{X}$ as:

$$\begin{aligned}\overline{X} &= 3\%(.12) + 5\%(.60) + 7\%(.23) + 9\%(.05)\\ &= (0.03)(0.12) + (0.05)(0.60) + (0.07)(0.23) + (0.09)(0.05)\\ &= 0.0542 \approx 5.4\%.\end{aligned}$$

   Notice that this average forecast is quite a bit higher than the actual inflation rate of 3.9%: the forecasters surveyed had a tendency to over-estimate inflation.

2. To calculate the standard deviation, we note first that the standard deviation is the square root of the variance, and second that the variance is roughly equal to the mean squared deviation (MSD). This approximation is better for larger n (sample size), but with "several dozen" economic forecasters in our sample, the approximation should not be too bad. (And this is the best we can do with the information we are given.)

   Using the formula for grouped data, the MSD is

$$\begin{aligned}MSD &= (3\% - 5.4\%)^2(.12) + (5\% - 5.4\%)^2(.60) + (7\% - 5.4\%)^2(.23) + (9\% - 5.4\%)^2(.05)\\ &= (0.000576)(0.12) + (0.000016)(0.60) + (0.000256)(0.23) + (0.001296)(0.05)\\ &= 0.0002024\end{aligned}$$

   Since the variance is approximately equal to the MSD, the standard deviation is approximately the square root of the MSD. Hence, $s = \sqrt{0.0002024} = 0.0142267354$, or 1.4%. Thus, a one-standard error confidence interval for the average inflation forecast is $5.4\% \pm 1.4\%$, or $(4.0\%, 6.8\%)$, which (barely) does not include the observed value of inflation. Given the magnitudes in question, this is a fairly large standard error — based on this survey, one would be (roughly) 95% sure that the actual 1983 inflation rate would lie between 2.6% and 8.2%. That's quite a difference!

2-41: 1. The average class size for an instructor is equal to

$$\frac{8+12+120}{3} = 46\frac{2}{3}.$$

2. 120 of the 140 students — or 6 in 7 students — views him- or herself as being in a very large class with 119 others. Certainly these students do not perceive the average class to be around 47 students. Thus to figure out the average class size from a student's perspective, we must compute a *weighted average,* with the weights corresponding to the relative proportion of students in each class.

   Thus, approximately 86% ($\frac{120}{140}$) of the students are in the largest class, while roughly 6% are in the 8 person class and 9% are in the 12 person class. (These proportions add to one, once rounding error is taken into account.) Thus the average class size for a student is

   $$\frac{120}{140}\cdot 120 + \frac{8}{140}\cdot 8 + \frac{12}{140}\cdot 12 = 104.3$$

   What explains the large difference between these two answers? Notice that the relative proportion of instructors assigned to each class is the same: 33%. So we implicitly computed a weighted average in part (a) above, but the weights were equal to $\frac{1}{3}$. In this problem, the weights are not equal, and the largest class gets far and away the most weight.