STAT 500 — HW 2 — Solution

All parts of Q1 and Q2 are one point each except for 2b that is 2 points. Q3 is 4 points. Q4 is 1 point.

1. Volatility on stock exchanges
   (a) No, this study is not a randomized experiment. We are observing the volatility in two major stock exchanges. Treatments (exchange) were not randomly assigned to stocks.
   (b) Yes, these data can be used to make inferences about the mean volatility of all stocks on the NYSE during the week of 12-16 Feb 2007. Because a random sample was selected from all stocks listed on the exchange that week. By randomly choosing stocks, the sample is representative (on average) of the whole exchange.
   (c)

$$S_p = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}} = \sqrt{\frac{(40-1)\times(2.23\%)^2 + (40-1)\times(3.37\%)^2}{40+40-2}} = 2.85\%$$

$$df = 40 + 40 - 2 = 78$$

   (d) $H_0: \mu_1 = \mu_2$ $H_a: \mu_1 \neq \mu_2$

$$T = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{S_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{2.91 - 4.39}{\sqrt{(2.85\%)^2\left(\frac{1}{40}+\frac{1}{40}\right)}} = -2.31$$

   $df = 78$, so p-value=0.024. Using a table, you should find $p < 0.05$. There is evidence of difference in mean volatility between the two exchanges.
   Note: Different tables have different d.f. That shouldn't change the conclusion.
   (e) No, I do not agree. Because the consultant's argument requires a causal conclusion. This study is not a randomized experiment.
   (f)

$$T = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{S_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{2.91 - 4.39}{\sqrt{(2.85\%)^2\left(\frac{1}{20}+\frac{1}{20}\right)}} = -1.64$$

   $df = 20 + 20 - 2 = 38$, so p-value=0.11.
   (g)

$$T = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{S_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{2.91 - 4.39}{\sqrt{(2.85\%)^2\left(\frac{1}{100}+\frac{1}{100}\right)}} = -3.66$$

   $df = 100 + 100 - 2 = 198$, so p-value=0.0003.
   (h) As the sample size increases, p-value becomes smaller and we are more likely to reject the null hypothesis. This is reasonable because a larger sample gives more information about the population which reduces the uncertainty about the mean volatility difference.
   Note: I asked the last three parts to emphasize that the p-value measures evidence against the null hypothesis. It is not a measure of whether the difference is important or large enough to care

about.
(i) 95% ci for the difference in mean volatility (NASDAQ-NYSE) is

$$\overline{x}_2 - \overline{x}_1 \pm t_{78,0.975}\sqrt{S_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} = 1.48\% \pm 1.99\sqrt{(2.85\%)^2\left(\frac{1}{40}+\frac{1}{40}\right)} = (0.21\%, 2.75\%).$$

(j)

$$\overline{x}_2 - \overline{x}_1 = 1.48$$

95% ci for the difference in mean volatility (NASDAQ-NYSE) is

$$\overline{x}_2 - \overline{x}_1 \pm t_{38,0.975}\sqrt{S_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} = 1.48\% \pm 2.02\sqrt{(2.85\%)^2\left(\frac{1}{20}+\frac{1}{20}\right)} = (-0.35\%, 3.31\%).$$

The ci becomes wider as the sample size decreases. Because the variability decreases as the sample size increases, we expect the ci to be more precise and narrower with large sample.

2. Sample size calculations
(a) $n = 40,\ \delta = 1.48\%,\ \sigma = S_p = 2.85\%,\ \alpha = 0.05$

$$\delta = (t_{2(n-1),1-\alpha/2} + t_{2(n-1),1-\beta})S_p\sqrt{2/n}$$

So $t_{78,1-\beta} = 0.324$. If you use the computer, you get power=$1-\beta = 63\%$. If you draw a picture, you see that the shifted t distribution is centered very near the dividing line between reject and not-reject. Hence the power is close to 50%. The t-quantile is > 0, so the power is $> 50\%$.
(b) $\alpha = 0.05,\ 1-\beta = 80\%,\ z_{0.975} = 1.96,\ z_{0.8} = 0.842$.

$$n \geq 2(z_{0.975} + z_{0.8})^2 S_p^2/\delta^2 = 58.2$$

Let n=59, then $df = 2(59-1) = 116,\ t_{0.975,116} = 1.98,\ t_{0.8,116} = 0.84$.

$$n \geq 2(t_{0.975,116} + t_{0.8,116})^2 S_p^2/\delta^2 = 59.2$$

Let n=60, then $df = 2(60-1) = 118,\ t_{0.975,118} = 1.98,\ t_{0.8,118} = 0.84$.

$$n \geq 2(t_{0.975,118} + t_{0.8,118})^2 S_p^2/\delta^2 = 59.2$$

Hence $n = 60$ in each group.
Note: Tables will usually have 100 or 120 d.f. I haven't checked, but both should give you a similar answer.
(c) $s.e. \leq 1.48\%,\ S_p = 2.85\%$

$$s.e. = S_p\sqrt{2/n}$$

$n \geq 7.4$, so $n = 8$ in each group.
(d)width$\leq 1.48\%$, $S_p = 2.85\%,\ z_{0.975} = 1.96$

$$\text{width} = 2t_{2(n-1),1-\alpha/2}S_p\sqrt{2/n}$$

Start with z to get $n \geq 113.96$.
Let n=114, then use $df = 2(114-1) = 226$, $t_{0.975,226} = 1.97$ to get $n \geq 115.2$.
Let n=116, then use $df = 2(116-1) = 230$, $t_{0.975,230} = 1.97$ to get $n \geq 115.2$.
Hence $n = 116$ in each group.

Note: The last three parts reinforce my point in lecture that the s.e. approach leads to the smallest n, the power approach for 80% (and also 95%) leads to intermediate n, and the ci width approach leads to the largest n.

3. Speed limits and traffic fatalities
The raw materials for my paragraph are:
the results from a t-test,
a 95% ci for the difference in means, and
the study design.

I got: mean difference = 22.3, s.e. = 7.3, 95% ci = (7.6, 37.0), p-value = 0.0038. All assuming equal variances and using the pooled s.d. Kelley's rule suggests that means should be reported only to the 1's place. I do that in my answer. This is an observational study, so I can't make causal conclusions. Notice that the boss's third question concerns the causal effect of changing the speed limit.

The sort of answer I expected from you is:

The change in traffic fatalities from 1995 to 1996 was analyzed using a t-test assuming equal variances. On average, the change in fatalities was 22% higher in states that increased their speed limits. The difference is not precisely estimated; a 95% ci for the difference is (7%, 37%), but there is strong evidence that it is different from 0 (p=0.0038). I can not predict the effect of changing the speed limit in our state without assuming that there are no other differences between the states that increased and the states that didn't.

Other points I would probably make, because I find them useful, are:
the means for the two groups, perhaps in a table.
the pooled s.d. (or perhaps the s.d. for each group).
the observation that results for individual states are highly variable (s.d. quantifies this, so does the min and max).
the states that increased their speed limits tended to be in the western half of the US, which is less densely populated.
You were not expect to provide these in your answer, because your boss didn't ask you for these.

Grading: total of 4 pts.
1 pt. Answer includes method.
1 pt. each Appropriate answers to each of the bosses questions
Deductions: -1 pt. values reported with more than 2 digits (e.g. 7.13%).
-1 pt. using unequal variance (Welch/Satterthwaite) test or interval
-1 pt. giving me more than one answer to each of the boss's questions.

4. The se (or ci) is not appropriate because we have enumerated the entire population of interest.

Notes: If you look at the description of the model-based approach, you see that we are using the data to make inferences about an unknown population mean. The se or ci describe our uncertainty in our conclusions about the population mean. What's the population here? It's the group of states. It's the same as the sample. We know the population mean for each group of states exactly.

From questions and comments, I suspect many answers say 'because you should use the pooled s.d. to construct the se for each group'. While I agree, there are huge differences of opinion here. Since I told you to assume equal variances, it would be better to use the pooled s.d. IF the sample was not an enumeration of the population. Because of that, this answer was worth 0.25 points.

Why is the s.e. of the difference reasonable? Not because of model-based inference, which has the same issue. But, model based inference is an approximation to design-based inference, which does make sense here. Is the observed difference large relative to might have been seen if the action was randomly assigned to states. Randomization tests condition on the observed data, so there is no problem if the sample is a complete enumeration.