1. LDL Cholesterol

(a) It is generally a good idea to include a placebo group to insure that any observed drops in LDL are due to the treatments and not to some other aspect of being in the study. For example, just being in an LDL study might encourage people to eat healthier. The data actually suggest that something like this may be happening as even the placebo group saw a drop in mean LDL.

(b) ANOVA

i. The relevant populations are individuals with high levels of LDL under the different treatment scenarios. Note that these “populations” refer to the same set of individuals (those with high levels of LDL) under a variety of treatment conditions. The population of interest is not the 100 people in the study. The parameters of interest is the summary of the population about which we want to draw inferences. In this case the parameters are the mean decreases in LDL after one year; the parameters are generally denoted with greek letters such as \( \mu_{\text{placebo}} \) for the placebo population.

ii. Using \( \mu \) (with appropriate subscripts) to denote the population means, the null hypothesis is that \( \mu_{\text{placebo}} = \mu_{a\text{--low}} = \mu_{a\text{--hi}} = \mu_{b\text{--low}} = \mu_{b\text{--hi}} \) and the alternative hypothesis is that the five means are not all equal to each other (i.e., at least two are different). It is not quite right to describe the alternative as \( \mu_{\text{placebo}} \neq \mu_{a\text{--low}} \neq \mu_{a\text{--hi}} \neq \mu_{b\text{--low}} \neq \mu_{b\text{--hi}} \) is a stronger statement than we want.

iii. SS(Within) = 19 \* 14.3^2 + 19 \* 12.6^2 + 19 \* 19.2^2 + 19 \* 12.9^2 + 19 \* 12.7^2 = 20132.2

SS(Betw) = 20*(10.1–18)^2 + 20*(11.2–18)^2 + 20*(19–18)^2 + 20*(23–18)^2 + 20*(26.7–18)^2 = 4206.8.

The F-statistic is \( \frac{4206.8/4}{20132.2/95} \) = 4.96 This should be compared to the \( F_{1.95} \) distribution which yields a p-value of .0011. If using the table, it is best to be conservative and compare to the \( F_{1.60} \) critical values which suggest that the p-value is between .001 and .005. The F-distribution for tests like these is always used in a one-sided manner. It is not appropriate to just arbitrarily pick a threshold for declaring significance unless you are asked to do so.

iv. The small p-value means that it is unlikely to observe a test-statistic as large or larger than the one we observed if the null hypothesis is true. This provides strong evidence that there are some differences among the population means.

(c) Contrast for drugs vs placebo

i. The estimated contrast is (0.25 \* 26.7 + 0.25 \* 23 + 0.25 \* 19 + 0.25 \* 11.2 – 10.1) = 9.875.

The estimated standard error of the contrast is \( \sqrt{\frac{4206.8/4}{20132.2/95}}(0.25^2(1/20) + 0.25^2(1/20) + 0.25^2(1/20) + 0.25^2(1/20) + 1/20) = 3.639. \)

Then a 95% CI is 9.875 \pm 1.985 \* 3.639 = (2.65, 17.10). If using the table, it is best to use \( t_{60,975} = 2.000 \).

ii. We are 95% confident that treating LDL with drugs leads to a mean decrease in LDL between 2.65 and 17.10 points larger than the decrease obtained with the placebo. Interpreting the confidence interval as a test, i.e., reporting that there is a significant difference, is not the correct answer.

(d) Drug A vs drug B

i. A natural contrast here is (0, –0.5, –0.5, 0.5, 0.5).

ii. This contrast is orthogonal to the previous contrast because the sample size in each group is the same and the sum of the products of the weights \((0 \* -1) + (-0.5 \* 0.25) + (-0.5 \* 0.25) + (0.5 \* 0.25) + (0.5 \* 0.25)) is zero.

iii. Orthogonality corresponds (under normality) to independence of the estimated contrasts.

(e) Power

i. If \( \alpha = .10 \) then our threshold for declaring statistical significance gets lower and we would need a smaller sample size.

ii. If we only expect a 15 point difference, then the effect size we are trying to detect is smaller and we would need a larger sample size to detect such a difference.

iii. If the within group s.d. is lower then the effect size is bigger (our measurements are more precise) and we would need a smaller sample size.
2. Faculty salaries

(a) Female vs male

i. This is a two sample confidence interval. Standard deviations are close enough that we can use the pooled procedure. 
\[ s_p^2 = \frac{(38 \times 26^2 + 357 \times 30.4^2)}{395} = 900.29 \]
\[ s_p = 30.0 \]
Then the 99% CI (with 395 d.f.) is
\[ 14.1 \pm 2.5883 \times 30.0 \times \sqrt{\frac{1}{39} + \frac{1}{358}} = (1.0, 27.2). \]
(Using the table would mean taking \( t_{120,995} = 2.617 \).) It is fine to use the unpooled procedure but then the degrees of freedom is not 395; if you can’t remember the formula (I can’t) then the safest thing to do is to use the minimum degrees of freedom (minimum sample size - 1 = 38). (The actual d.f. in this case is about 50.)

ii. The point of this question is to note that this is an observational data and the two groups may differ in many other ways. This type of analysis can’t prove gender discrimination. Many people noted that statistics can not prove anything with certainty; this is true but not the primary point in my view.

(b) Years of service regression

i. The estimate and standard error are given. The 95% CI is
\[ 0.7796 \pm 1.966 \times 0.1104 = (0.56, 1.00). \]

ii. There is evidence of a quadratic shape in the plot which suggests the relationship may not be linear in years of service. Also evidence of increasing variance. No major issues are evident in the normal probability plot – there are two standardized residuals bigger than 3 when only one would ordinarily be expected in a sample of this size. I looked for people to address these three assumptions. I generally accepted any interpretation but did deduct a small amount for missing features that I thought were obvious.

iii. If you recall that the prediction standard error is
\[ s_e \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}} \]
then you could calculate that the width of the interval is approximately
\[ 4s_e \] (this is \( \pm 2 \) standard errors) which is approximately 114. Thus answer is that the width is closest to 100. Note even if you didn’t remember that you should always remember that the \( s_e \) is the quantity that measures expected variation around the regression line which should lead you to the same logic.

iv. You knew this question was coming. The probability of obtaining an estimated slope (or test statistic) like the one we observed or a value more extreme if there is in fact no linear relationship between the variables is .04.

(c) Multiple regression

i. The intercept here is the expected salary for the subpopulation of professors with all predictors equal to zero. In this case that means the subpopulation of male assistant professors with zero years of service.

ii. The coefficient of YrsService is the change in the expected salary for a one-year increase in YrsService with all other variables held fixed.

iii. The estimated coefficient of female here gives the difference in the expected salary between two subpopulations, one consisting of male faculty and one consisting of female faculty where both subpopulations have the same number of years of service and are at the same level of appointment. The earlier difference did not control for years of service or faculty rank - we can conclude that there are differences between males and females with respect to these other variables. In fact, females have less service on average (11.5 yrs) than males (18.3 yrs) and this explains part of the salary difference that we saw in the two sample analysis of part (a).