INSTRUCTOR NOTES FOR M358K FOR PART VI (LEARNING ABOUT THE WORLD: CHAPTERS 23 - 26) OF DEVEAUX, VELLEMAN AND BOCK, *STATS: DATA AND MODELS*, 3RD ED.

Chapter 23: Inferences About Means

- pp. 550-551 The example doesn't seem to be the best choice for testing a mean – it would make more sense to look at the proportion of cars exceeding the speed limit rather than looing at the mean speed.
- p. 551 The statement (second paragraph of Getting Started), "We always center confidence intervals at our best guess of the unknown parameter," is a little misleading: The techniques in this textbook do that, but some more advanced techniques (which rely on skewed distributions) do not.
- p. 552 The For Example and following discussion seem well done. However, the ActivStats activity Estimate the Standard Error didn't function properly when I tried it. (It wouldn't let me enter the correct answer!)
- pp. 553 – 556
    - Supplement these pages with the handout *Chi-squared Distributions, t-Distributions, and Degrees of Freedom*.
    - p. 553 The comments below and next to Gosset's picture are interesting. But the "Notation Alert" neglects to mention the possible confusion from the fact that t is also often used for time.
    - p. 554
        - The last sentence in the For Example needs amendment to take into account that outliers from one farm were removed. Of course, the outliers could have been just recording errors, but in case they were not, it would be better to say something like, "farm-raised salmon from typical farms, " and to present the analysis with the problem farm included as well.
        - The ActivStats activity Learn About Student's Distributions might be useful for interactively showing the effect of degrees of freedom on the shape of the t-distribution (although you need to let it run about three-quarters of the way through to get to the place where the slidebar allows you to do this).
    - p. 555 The blue box says it well. Note that the correction formula mentioned at the bottom of the page was given in the supplement *Where Does the 10% Condition Come From?* for Chapter 18.
    - p. 556 Very good – including the blue box and the footnote.
- p. 557 The Just Checking is very good.
- pp. 557 – 558 The Step-by-Step Example (referring to the situation on p. 550) makes some good points, especially the Reality Checks and the Caveat.
- p. 559: The Cautions section is good. (Be sure to assign Exercises 5-8 to reinforce these points.)
    - I have found web demos such as the following to be useful in helping students understand confidence intervals (My preference is the Freeman demo):
        - Bioconsulting Confidence Interval Simulation

        - [R. Webster's Confidence Interval Simulation]
        - [W. H Freeman's Confidence Interval Simulation]
        - [The Rice Virtual Lab in Statistics Confidence Interval Simulation]
    - Comments on relevant ActivStats activites:
        - "Gain Intuition for t-based Intervals" might be helpful for reinforcing the importance of and difficulties in checking model assumptions.
        - "Simulate Intervals for Small Samples" tries to make a nice point, but the simulation graphs used don't seem the best for showing the point.
        - "Compute a Confidence Interval" might be helpful for students who are still struggling with understanding confidence intervals – it deals with a large sample, so approximating sigma by the sample standard deviation and using a z-model is reasonable. But students may need to have it pointed out that this simplification is only appropriate when the sample size is large.
- p.p. 560 – 562 are well done.
    - The advice in "Make a Picture" promotes good statistical practice.
    - The For Example and Step-By-Step Example model good statistical practice.
    - The ActivStats activity Simulate to Test the Mean of an Unknown Population seems to be a good idea that wasn't well enough implemented to do what was hoped; I'd be reluctant to use it without some further explanation.
    - The ActivStats activity Learn to Use the t-tables might be interesting for some students (and possibly worthwhile for students who would rather use a table than a computer) – it really discusses the limitations of t-tables more than how to use them.
- p. 563: The Just Checking is nice.
- p. 564-565:
    - The ActivStats activity Understand How Tests and Intervals Are Related might be helpful for a student who doesn't get the relationship but who is comfortable working with the format used in the ActivStats activities.
    - The ActivStats activity Learn the Real Effect of Small Sample Sizes does not seem to do what it purports to do (at least not well) – the instructions were not clear to me (what is the red region and what is the marked region?), and there were some glitches in operating it.
    - The discussion of Sample Size is well thought out.
- p. 566 An interesting heuristic for degrees of freedom. But see the handout *Chi-squared Distributions, t-Distributions, and Degrees of Freedom* for a more mathy discussion.
- pp. 566 – 568 The Sign Test is optional – fine to cover if you have some extra time. Or just refer students to it if it is appropriate for their project. The discussion following the example is good – be sure to read it yourself even if you don't cover the sign test.

- pp. 568 – 570 As usual, the What Can Go Wrong is good. The Connections section is also particularly worthwhile.
- pp. 571 – 572 Inference for Means on the Computer
    - The comments on p. 571 are particularly worthwhile.
    - p. 572 is worthwhile for students to read.
- Exercises (pp. 574 – 579 and supplement):
    - #5 - 8, 25 – 28, and 37 – 39 are strongly recommended.
    - Other suggestions for exercises from the textbook: 3, 4, 9-18, 29, 30, 33, 34
    - Exercise 1 from the handout *Chi-squared Distributions, t-Distributions, and Degrees of Freedom* is recommended.
        - If you have time, Exercise 2 from the handout would be good to reinforce the definition of chi-squared distribution, as well as properties of the mean and variance.
        - You might also consider having students do their own simulations like the one discussed in item (i) on p. 8 of the handout – preferably with each student using a different choice of distribution, and then comparing results in class.

Chapter 24: Comparing Means

- p. 581 "Plot the Data" gives a good, realistic discussion of outliers.
- p. 582 You may want to supplement the discussion (including the footnote) of the t-test for the difference of two means by giving the formula for the pdf of a t-distribution, and pointing out that it makes sense even if ν is not an integer:

  $$f(t) = \frac{1}{\sqrt{\pi\nu}}\frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)}\frac{1}{(1+t^2/\nu)^{(\nu+1)/2}}, \text{ for } -\infty < t < \infty,$$

  where ν = degrees of freedom and Γ denotes the gamma function.
  (But note that the students might not be familiar with the gamma function)
- p. 583 The ActivStats activity Learn the Assumptions of the Two-Sample t-Test might be helpful for students who find the type of format it uses helpful.
- p. 584
    - The ActivStats activity "Compute a Confidence Interval for the Restricting Diet Data" worked well when I tried it. It might be worth assigning, suggesting, or going over in class.
        - Note that the activity refers to the situation described in the ActivStats Video about Diet Prolonging Life, so needs to be preceded by the (very short) video
        - But be aware that the quiz on the video didn't seem to be functional for me.
        - One reason to include this would be that additional research on the subject of calorie restriction has recently been published. See, for example, Gina Kolata's NY Times article http://www.nytimes.com/2012/08/30/science/low-calorie-diet-doesnt-prolong-life-study-of-monkeys-finds.html?_r=0. This points out the importance of replication and the many

problematical aspects of performing scientific studies and analyzing them statistically.
    - The point in the blue box is worth pointing out.
- pp. 585-587 The examples are well done, especially the Reality Check, the model checking, and the comments on choice of degrees of freedom. The Just Checking has some good questions.
- p. 588 The ActivStats activity might be helpful for students who need to see additional examples. They might want to view the brief video on Hot Dog Nutrition first.
- pp. 589 – 591 The Step-By-Step Example is especially good, particularly the discussion of model assumptions and even more so the "Reality Check" (top of p. 590) and the caution at the very end of the left column on p. 591.
- p. 591 Question #7 of the Just Checking is particularly good.
- p. 592 In the For Example, note the openness about uncertainty in the model assumptions – and that the conclusion says, "provides evidence" rather than "proves". These both model good statistical practice.
- pp. 592 – 596
    - Back Into the Pool gives a nice introduction to the Pooled t-Test.
    - Emphasize the last sentence before The Pooled t-Test and the sentence at the bottom of p. 593
    - Also emphasize "Is the Pool All Wet?" and "Why Not Test the Assumption That the Variances Are Equal."
    - Point out in connection with "Is There Ever a Time …" that having different variances is indeed a difference between distributions for two groups – and sometimes a very important difference. Examples:
        - A difference in variances of temperatures in two locations could make a big difference in survival of a species in those two areas. For example, a species that is especially sensitive to high (or low) temperatures could have lower survival in a region with high temperature variance than in a region with less variable temperatures, even if mean temperature in both locations were the same.
        - Since both high and low blood sugar can have negative effects, high variance of blood sugar in patients taking a medication would produce a health hazard not present in a medication that produced the same mean blood sugar but had variance low enough to keep blood sugar within a healthy range.
    - Emphasize the question mark by the Nearly Normal Condition and the comment that "The presence of skewness and outliers in groups this size (27) suggests that using t-methods may be risky."
- p. 597 Omit the starred sections unless you have extra time. But refer students to them if their projects turn out to require them.
- p. 598 Both points in What Can Go Wrong are important – and the box following makes an important point.
- p. 599 The Connections box does a good job of giving the big picture of what has been studied so far.

- pp. 601 – 602 The general advice (box from p. 601 through top of p. 602) is good – be sure to have students read it before using software.
- Exercises (pp. 603 – 610):
    - Strongly recommended: #5, 6, 13, 16, 21, 31
    - Other suggestions: #1-4, 7, 9 – 12; parts (a) –(c) only of 14; maybe 22 if 21 is also assigned.
    - Not recommended:
        - #8 (The wording seems iffy in a couple of places)
        - #27, 28 (Because they say "Assume any assumptions and conditions that you cannot test …")
        - #16, 23-26, 30, 33 – 34, 36 – 38 (Because they deal with three groups using two-group methods, and might mislead students into doing this, with no discussion of why ANOVA might be preferable nor of overall type I error rate and the effect of multiple inference on it.)

Chapter 25: Paired Samples and Blocks

- This chapter is for the most part especially clearly and well done, so just a few comments.
- p. 613 The ActivStats activity Compare the Difference in Means of Paired Groups might be helpful for students who learn well from the format of these activities.
- p. 614 The discussion of the 10% condition seems a little confusing – one *does* need to check the condition for an observational study.
- p. 615 Emphasize the comment that it is not legitimate to use the data to decide which type of alternate hypothesis to use.
- p. 617 The Reality Check at the very top is a good example of statistical reasoning. Emphasize the Note at the bottom of the page.
- p. 620 The discussion of Effect Size is good, although it may take a little thinking to see how the graph shown supports the point being made.
- p. 621
    - The middle paragraph under Blocking is important – it illustrates how spurious reasoning based on incomplete knowledge can lead to inappropriate statistical analyses.
    - The last paragraph under Blocking is also important.
    - The Just Checking includes good questions that address common points of confusion.
- pp. 621 – 622 The Sign Test is optional – cover it only if you have extra time. Refer students to the section if it is appropriate for their project.
- p. 623 What Can Go Wrong addresses common misconceptions. "Connections" is good.
- pp. 624 – 625 Paired t on the Computer: The general discussion (p. 625 – top of p. 625) is good; it would be good to assign for reading. Most students probably need an explanation of (or at least need to think about) the comment , "… a paired t-test is about whether the points tend to be above or below the line y = x," on p. 625.
- Exercises (pp. 626 – 632):

  - Suggestions to choose from: #1, 2, 5 – 15,18, 19(a) – (c), 20, 21, 22(a) – (d), 25, 27(a) and (b), 29 – 32
  - There appears to be a misprint in #6 (presumably Cloud , not Could)
  - It's not clear to me what the numbers in the table in #17 are supposed to stand for.

Chapter 26: Comparing Counts

- For the most part, I find this chapter explains the various chi-square tests better than any other text I've encountered.
- The Activ/Stats activities for this chapter don't seem to correspond well with this textbook. Among other things, they don't fit the progression from goodness-of-fit-test to test-of-homogeneity to test-of-independence, which seems to be one of the strong points of the chapter.
- pp. 633 – 641 (Goodness of Fit test) Starting with the goodness-of-fit test seems to be a good idea, rather than getting into the subtle distinction between a test for homogeneity and a test for independence at the same time as introducing the terminology and somewhat messy formulas.
  - pp. 624 – 635 (Assumptions and Conditions)
    - Stress the "Counted Data Condition." It often goes unemphasized, leading to the misuses mentioned on p. 634.
    - Note that the "10% Condition" is (presumably inadvertently) not explicitly mentioned in the list of Assumptions and Conditions – but it is mentioned in the For Example. If dealing with a finite population (especially when sampling without replacement), it is relevant.
  - pp. 635 – 639 (Calculations)
    - Pointing out the connection with residuals used earlier seems like a good idea not often mentioned in texts.
    - Some students may find an explicit example helpful in explaining why the squared deviations are divided by the expected counts. Here's a possibility: Suppose one category has expected count 100 and observed count 95, while a second category has expected count 10 and observed count 5. Then the observed count is 95% of the expected count for category 1 – pretty close. But the observed count is only 50% of the expected count for category 2 – pretty far off. So we'd want the far-off count in category 2 to count more than the close count in category 1. But both categories have difference Obs – Exp = 5. That's not good. However, if we divide the squared deviation by the expected count, category 1 contributes 25/100 = .25 to the chi-squared statistic, whereas category 2 contributed 25/10 = 2.5, giving more importance to category 2 in the chi-squared statistic, as we would hope.
    - Emphasize the point on p. 636 that n is the number of categories, not the sample size.
    - You may need to point out that since the chi-square statistic is the sum of squares, it must be ≥ 0. Also, the formula for the statistic

shows that the farther the observed values are from the expected values, the larger the statistic. Thus, the test is necessarily one-sided.

- Point out that, like the z-tests for proportions, the chi-squared test is only approximate.
- You may want to remind them of the definition of chi-squared distribution in the handout Chi-squared Distributions, t-Distributions, and Degrees of Freedom (supplement to Chapter 23)
- You may also want to have them read, or go through in class, the handout Connection Between the One-Sample Test for Proportions and the Chi-Squared Goodness-of-Fit Test .
- I don't have students do complete chi-squared calculations on exams, but will often include a question where I give computer output with some parts deleted and have them fill in the missing parts, to check if they understand the formulas and reasoning.

- p. 640 The box has some nice mathy stuff
- pp. 640 – 641 "But I Believe the Model …" makes an important point for application.

- pp. 641 – 647 (Chi-Square Test for Homogeneity)
  - p. 641 See the supplement Connection Between the Z-Test for Proportions and the Chi-Square Test of Homogeneity for more on the comment (very end of page) that "The z-test for two proportions generalizes to a chi-square test of homogeneity." Students may need to read through the top of p. 645 before reading the entire handout. (On the other hand, some of them might need to read the handout before understanding the discussion of calculations in the text. See also the next comment)
  - p. 642
    - I'm not convinced that the explanation of when it's appropriate to do a chi-square test of homogeneity is as good as it might be. I've tried to give a summary on the handout mentioned above. See what you think.
    - Rice (*Mathematical Statistics and Data Analysis*, 3rd ed, 2007, p. 516) describes the chi-square test of homogeneity as applying to a situation where we have independent observations from J multinomial distributions, each with I cells; we want to test whether the probabilities for the two multinomial distributions are the same. He also points out (p. 523) that the chi-square test for homogeneity is derived based on the assumption that the column (or row) totals are fixed.
  - pp. 642 – 643 For this class, go through (or have them go through – or maybe better yet, have them both guess and explain) the general formula:

$$\text{Expected count for cell in row i and column j} = \frac{(sum\ of\ row\ i)(sum\ of\ column\ j)}{sum\ of\ all\ counts\ in\ table},$$

or for short:

$$\text{Exp} = \frac{(row\ sum)(column\ sum)}{grand\ total}$$

- p. 644
  - Nice discussions of suitable graphs and of model assumptions.
  - Note the table including both observed and expected counts. This is a variant of what software output often includes: a table with both observed and expected counts in each cell.
- p. 645-646
  - I haven't encountered standardized residuals for chi-squared tests before. I haven't been able to find or figure out a good explanation of why the variance of the expected value should be the expected value; the closest I've found is that Pearson introduced these residuals (often called Person residuals), basing them on the assumption that the counts have a Poisson distribution, hence have variance equal to expected value. (But why should the counts have a Poisson distribution??) However, the notion does give a handy way of thinking about the further analysis. If the iffiness of standardized residuals bothers you (as it does me), you can fall back on the fairly common notion of "chi-squared components," which are just the squares of the Pearson residuals. Since they are what is added up to make up the chi-square statistic, saying that cells with larger contributions to the chi-square statistic indicate groups that are most likely to fall in a particular category. *However*, the standardized residuals do have the advantage that their sign (positive or negative) has an interpretation. (More on this on pp. 650 ff.)
  - Some people may be tempted to follow up a chi-square test that supports a difference by doing more hypothesis tests on smaller groupings. This is *not* good practice unless you account for multiple comparisons. (See pp. 733 – 735 for a discussion of this in the context of Analysis of Variance, and http://www.ma.utexas.edu/users/mks/statmistakes/multipleinference.html and/or http://en.wikipedia.org/wiki/Multiple_comparisons for more discussion and references.)
- pp. 646 – 647 #6 of the Just Checking is nice.

- pp. 647 - 652 (Chi-Square Test for Independence)
  - The tattoo example is a good choice, since it provides an opportunity to discuss what might be done when expected cell counts are too small.
  - p. 647 The blue box at the bottom of the page is a good, pithy reminder of how to tell the difference between the chi-square tests of homogeneity and independence. Adding to the subtleties: Rice (*Mathematical Statistics and Data Analysis*, 3rd ed, 2007, p. 523) points out that the chi-square test for independence is derived based on the assumption that only the overall total of counts is fixed -- in contrast to the assumption of the chi-square test for homogeneity that the column (or row) totals are fixed. He also points out that independence can be thought of as homogeneity of conditional distributions.

    - pp. 648 – 650 Be sure to draw students' attention to the cautions about low expected counts in the Step-By-Step Example.
    - p. 650 Illustrating a point mentioned above: In the "For Example," it may be tempting to calculate the residuals by taking the square roots of the chi-square components - - for example, calculating the standardized residual for the cell in the example as "standardized residual = $\sqrt{14.75}$ = 3.84. However, this would be wrong if the observed value for the cell were smaller than the expected value – the standardized residual would in that case be -3.84, not 3.84. Note that the sign *is* relevant for the type of analysis done for the tattoo example on p.651.
    - p.651 Emphasize the following points:
        - (middle of page) "Too small an expected frequency …"
        - The use of "*suggest"* in the last sentence on the page – this illustrates good practice.
    - p. 652
        - Emphasize the section "Chi-Square and Causation."
        - The "Just Checking" has well-chosen examples. (I had to think carefully, but got them all correct when I did. :~))
    - p. 653
        - As usual, the WCGW is important. Be sure to emphasize the second point, "Beware large samples." For any hypothesis test, large samples may "detect" a statistically significant difference that is not practically significant. Forgetting the difference between "statistically significant" and "practically significant" is all too common (both are important!), but with chi-square tests, the lack of effect size makes it even easier to neglect practical significance and focus (possibly misleadingly) on just statistical significance.
        - Related to the preceding point: You might hear it said that for chi-square tests, "the test statistic grows proportionally to sample size." I think this refers to the following observation: If you multiply every count in the contingency table by a constant c, then the chi-square statistic is multiplied by c. Here is the argument in words: Multiplying every count by c multiplies each row sum, each column sum, and the grand total by c. Recalling that expected count = (row sum)(column sum)/grand total, we see that each expected count is multiplied by c as well. It follows that each chi-square component (hence the chi-square statistic) is multiplied by c.
        - Emphasize the "not" in the second paragraph of the Connections box.
- p. 655: The top (general) part of "Chi-Square on the Computer" is worth having students read.
- Exercises (pp. 657 – 663) Suggestions to choose from:
    - Both of #1,2
    - Self-check: #3, 5, 9; odds from 13 – 21 as one problem; 27, 29, 31; 33 and 35 as one problem.

- Hand in: Some selection from #4, 8, 10; evens from #14 – 22 as one problem; 28, 30, 32
- Note: I've tried to leave out exercises including ordinal data, since they might be better analyzed by methods using the ordering in the data.
- Idea for an additional problem (adapted from Rice, p. 540): Chi-square tests for homogeneity and independence calculate the test statistic from counts. What would happen if you tried to calculate the chi-square statistics from proportions instead of counts? If you tried to calculate it from percentages instead of counts?

Part VI Review (pp. 664 – 665): I strongly recommend that students go over this before trying the review exercises (to help see both the big picture and the details), and consult it when doing the review exercises and studying for exams.

Review Exercises (pp. 665 - 672:

- Suggestions to choose from: #3, 4 (parts (c) and (d) are especially good), 5, 6, 8, 9, 10, 11, 12 (especially the nice range of questions), 18, 19, 21 (although the first sentence seems irrelevant), 23, 24, 26, 27, 29, 31 (a nice variety of questions), 32? ((c) is stated not too clearly), 33? (poor wording in the first sentence), 34, 35, 36 (a nice combination of questions), 37, 38, 39 ((a) and (e) are nice; I might quibble about (c), since details of selection are not given), 40, 41, 42
- Comments on some I don't recommend: 1 -- the answer rejects with p-value 0.059, without giving any justification; 2 -- statement is ambiguous: time vs. ability; 7 – I'm not convinced that the test used is the right one for the problems statement; 13 – neglects the problem of multiple testing; 14 – I'm not convinced that the supporting reasons are the best; 15 and 16 –the question asked doesn't fit the data well; 17 – (b) is poorly worded; 20 – poorly worded; 22 – There doesn't seem to be a justification for concluding "European …" when three countries are excluded; 25, 28 and 30 – neglect the problem of multiple inference.