MS Module 7 Hypothesis testing – *p* values (overview 3rd edition)

(The attached PDF file has better formatting.)

(Readings from the third 3rd edition of the Devore, Berk, and Carlton text.)

Two complex statistical procedures that are not use in actuarial work (Levene's test and Tukey's procedure) have been removed from the syllabus, simplifying modules 10-14 on analysis of variance (ANOVA). To keep the 24 module sequence,

- Module 4 Hypotheses and Test Procedures is now split into
  - Module 4a Type 1 and Type 2 errors:
  - Module 4b Tests about a population mean

- Module 5 Hypothesis testing of proportions is now split into
  - Module 5a Tests About a Population Proportion
  - Module 5b Hypothesis testing – *p* values

Reading: §9.4: *p* values
Reading: §9.5: Comments on Selecting a Test Procedure, subsections
- statistical versus practical significance
- power and uniformly most powerful tests (through the end of example 9.22, including Figure 9.10)

The other subsections in §9.5 ("best tests for simple hypotheses" and "likelihood ratio tests") are not on the syllabus for this course.

Example 9.18 illustrates the meaning of the *p* value. Social scientists often say that an experiment provides no evidence for a hypothesis; the better statement is that the *p* value exceeds some critical value. These critical values are (to some degree) arbitrary; no scientific reason exists for a critical value of 5% or 1% instead of 6% and 2%. The proper statistical procedure is to show the *p* value; the reader can then decide whether the *p* value is low enough to doubt the null hypothesis.

In genome-wide association studies (GWAS), a million independent hypotheses may be tested, so the critical *P*-value is generally set at $5 \times 10^{-8} = 0.00000005$, so as to give a 5% expected number of false positives. This procedure makes sense, even if it is not entirely persuasive. But if the computed *p* value for a particular result is 0.00000006, one would not say that the experiment provides no evidence. Geneticists debate what critical *p* value is proper. A value of 1% would give 10,000 false positives per experiment, but a value of 0.000005% may give tens of thousands of false negatives per experiment.

Some geneticists are perplexed by the *p* values. They may consider only a hundred hypotheses to be likely, but to save research costs, they test another million. (Most genomic research costs are per experiment, not per hypothesis.) Had they tested only the first hundred, all results would be highly significant; why should the next million independent hypotheses affect the significance of the first hundred? This argument is somewhat exaggerated, since geneticists usually don't know which hypotheses are most likely, but it captures the current controversy over *p* values.

Subjective statements, such as "the data do not support the view that …," are often invalid. The effect size may be large, the sample size small, and the *p* value may be 2%, but the researcher was using a 1% critical *p* value. Conversely, the effect size may be small, the sample size large, and the *p* value may be 9%, but the researcher was using a 10% critical *p* value. In practice, many non-mathematical readers want binary results: something is true or false, right or wrong. Hypothesis testing gives probabilistic statements amid a host of qualifying attributes: *p* values along with sample size, effect size, conditioning of other variables, etc.

To save time, memorize the most common critical p values: $z_{0.10}$, $z_{0.05}$, $z_{0.01}$, for both one-tailed and two-tailed tests. For small or moderate samples with a t distribution, you must interpolate from tables or use Excel (or other statistical software).

Excel has built-in functions for computing p values. When you work the problems in the textbook or on the discussion forum, check your results with Excel.

Final exam problems often ask for the p values (by interpolation in the tables provided). The exam problems are multiple choice questions, and the choices are sufficiently different that simple interpolation is sufficient.

Know well the sub-section "*P*-Values for *z* Tests." This section is simple, and it is the foundation for the more complex relations in later modules. The sub-section "*P*-Values for *t* Tests," is similar, though it has an additional parameter (the degrees of freedom) and more diffuse curves. Know examples 9.20 and 9.21.

Example 9.22 is illuminating and worth reading, but you will not be tested on the shapes of these histograms.

The final exam does not test the proof of the Neyman–Pearson theorem. You may skip Example 9.23; the mathematics in this example is not tested on the final exam.

Regulators and journals set critical z values: new drugs must meet strict thresholds to be deemed safe and journal submissions must meet thresholds to be published. These critical z values and critical t values (such as 5% or 1%) are arbitrary. They help standardize regulatory tests and papers in research journals, but they are not appropriate for business decisions.

Actuaries are concerned with both p values and the economic costs of using new research. A p value of 6% when the costs of using the new information are small is better than a p value of 4% when the costs are high.

The textbook refers to these costs as statistical versus practical significance. These costs differ for each use of the statistical analysis; no general formulas can be given.

The textbook shows how p values are computed. Final exam problems ask for p values, which you derive by interpolation on the exam. (In practice, computer programs give the p values for most distributions).

Review end of chapter exercises 49, 50, 51, 52, 53 (for a one-tailed test, you must check the sign of t), 54, 55, 56, 57, 58, 59, and 61.

Section 9.5 "The Neyman–Pearson Lemma and Likelihood Ratio Tests" is not on the syllabus for this course. Likelihood ratio tests are becoming more widely used, especially since R simplifies their calculation, so parts of this section are worth reading, but this section is not tested on the final exam.