MS Module 8: Inferences Based on Two Samples (overview 3rd edition)

(The attached PDF file has better formatting.)

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

Reading: §10.1: The Two-Sample *z* Confidence Interval and Test.

Much actuarial pricing and risk classification deals with means of two or more groups of policyholders, such as men vs women or rural residents vs urban residents. The *Z* statistic is the difference in the sample means divided by the standard deviation. The variance of the difference of two independent random variables is the sum of their variances, and the standard deviation is the square root of this variance. For the variance of the difference in means, each variance is divided by the size of its sample.

If you lack a statistics background, review

§1.4 Measures of variability (especially σ, σ^2)
§5.2 Covariance and correlation
§5.4 Transformations of random variables
§6.2 Distribution of the sample mean (especially the central limit theorem and the law of large numbers)
§6.3 Distribution of a linear combinations (especially of differences between random variables)

Know the variance of combinations of random variables, especially of the difference of two independent random variables. Equation 10.1 is used throughout this chapter. Example 10.1 is simple, but it illustrates the concepts well. Example 10.2 shows how the procedures are used, following the seven step sequence. Final exam problems ask you to derive *t* values (steps 4 and 6), *p* values, required sample sizes, and α and β for type 1 and type 2 errors. Know the three variants shown directly above Example 3 of the equation for β, and the formula for the required sample size to achieve selected values of α and β.

The null hypothesis is that the difference in means is some pre-determined value, not necessarily zero. An actuary may test a hypothesis that the expected life is K years longer for people of Country S vs Country T.

Causation is hard to infer from observational studies. Residents of a country differ many ways: income, health care, education, and diet are examples that may affect mortality rates. The textbook discusses inferences of causation several times; the modules on two-factor ANOVA and insurance risk classification explain how the inferences are interpreted.

Causation is controversial for actuarial pricing and risk classification. Urban residents in developed countries have higher motor insurance accident frequencies and pay higher rates than rural residents pay. Residence does not cause accidents, and an urban resident who moves to a rural area is charged a lower rate. The two views to this controversy are

- Accident frequencies depend on characteristics of the policyholders, and residence is the best available predictor of these characteristics.
- Residence over-charges some policyholders and under-charges others, so actuaries should focus on the true characteristics affecting accident frequencies.

A similar controversy affects mortality rates by country or region. Insurers in east Asia assume lower mortality than insurers in other countries, but a person of east Asian extraction pays the same rate as other consumers for insurance contracts bought in other countries.

Review end of chapter exercises 1 a and b, 2, 3, 4, 6, 8, and 9.

The textbook gives formulas for β (the probability of a Type II error) for three types of alternative hypothesis: more than, less than, and not equal to the hypothesized difference in means. For each formula, know the area

under the normal curve to which it refers. Recalling the abstract formulas is difficult, since the parameters depend on the scenario in the section of the text, but the graphic remains the same. The textbook shows the graphics for several of the scenarios.

If the sample is large enough, the population need not be normally distributed and the sample variance S^2 may be used in place of the population variance σ^2. This application of the central limit theorem is used repeatedly in the textbook.