Notes for M 384E, Wednesday, January 19, 2005

This is an introductory course in Analysis of Variance and Design of Experiments.

I. QUESTIONS:

1. What is Analysis of Variance?
2. How is Analysis of Variance connected to Design of Experiments?

II. BRIEF ANSWERS:

1. Analysis of Variance (ANOVA) is a methodology that can be used for statistical inference in a variety of situations generalizing the equal-variance two-sample t-test.

2. The details of implementation of ANOVA depend on the design of the method for collecting data -- typically, by an experiment. The design needs to take into account the methods of analysis as well as the particulars of the context (the question of interest, factors that may influence the variable of interest, and constrains such as time and budget.)

III. REVIEW OF THE EQUAL VARIANCE TWO-SAMPLE T-TEST (focusing on the model assumptions and why they are important.)

(There is another two-sample t-test that does not assume equal variance. However, the equal variance test is the one we will be concerned with here.)

Be sure to review the handout Review of Basic Statistical Concepts if you are confused about notation or basic concepts used below.

Statistical procedures typically have certain *model assumptions*. These are assumptions about the distributions of random variables involved, or about how samples are chosen, or about the type of relationship between the variable involved. The essence of applying statistics is to find a model that does all of the following:

1. Fits the real-word situation involved.
2. Leads to a valid method of statistical analysis.
3. Gives information relevant to the questions of interest.

In using statistics, always bear in mind the words of the statistician G.E. Box:

All models are wrong; some are useful.

This means that models never fit the real-world situation exactly, but we need to be sure that they fit "well enough" and that they give relevant information.

Model Assumptions for the equal variance two-sample t- test:

1. $x_1, x_2, \ldots, x_m$ and $y_1, y_2, \ldots y_n$ are *independent, random* samples from random variables X and Y.
2. X and Y are each *normally distributed*.
3. X and Y have the *same variance* (which is not known)

Denote the means of X and Y by μ_X and μ_Y, respectively.

We wish to test the null hypothesis $H_0: \mu_X = \mu_Y$
against the two-sided alternative $H_a: \mu_X \neq \mu_Y$

Example: A large company is planning to purchase a large quantity of computer packages designed to teach a new programming language. A consultant claims that the two packages are equal in effectiveness. To test this claim, the company randomly selects 60 engineers and randomly assigns 30 to use the first package and 30 to use the second package. Each engineer is given a standardized test of programming skill after completing the training with the assigned package. The scores of the 30 engineers assigned to the first package are $x_1, x_2, \ldots, x_m$; the scores of those assigned to the second package are $y_1, y_2, \ldots y_n$. (In this example, n = m = 30.) The random variable X is "test score of someone using the first package." The random variable Y is "test score of someone using the second package". Since the engineers are randomly chosen and randomly assigned to the package, assumption (1) is satisfied. Since the test, like most standardized tests, is devised and scored to have a normal distribution of scores, assumption (2) is reasonable. It is reasonable to assume that the variability in scores will not depend on the package chosen, so assumption (3) seems reasonable (although perhaps we might want to look the data to get an additional check on whether this assumption is reasonable).

Outline of what the test involves and why it works (focusing on where the model assumptions are needed):
(For more details, see Ross, Section 4.2 or Wackerly Section 10.8)

Denote the (unknown) variance of X and Y by σ^2.

Notation: The notation $X \sim N(\mu_X, \sigma^2)$ is sort for "the random variable X is normally distributed with mean μ_X and variance σ^2". Thus from our assumptions:

$$X \sim N(\mu_X, \sigma^2) \text{ and } Y \sim N(\mu_Y, \sigma^2)$$

From our sample $y_1, y_2, \ldots y_n$, we can calculate the sample mean $\bar{y}$, which is our best estimate of the mean μ_Y. We could also calculate the sample mean for *any* random sample of size n chosen from Y. This process ("take a random sample from Y and calculate its sample mean") describes a new random variable, which we will call $\overline{Y}$. Thus, $\bar{y}$ is the value of the random variable $\overline{Y}$ obtained by picking our particular sample. Since $\overline{Y}$ is a random variable, it has a distribution (called a *sampling distribution,* since the value of $\overline{Y}$ depends on the sample chosen) . Mathematical theory tells us that the random variable $\overline{Y}$ is normally distributed with mean μ_X and variance σ^2/n: $\overline{Y} \sim N(\mu_Y, \sigma^2/n)$

This conclusion uses the facts that we have <u>random</u> samples and that $Y \sim N(\mu_Y, \sigma^2)$.

Similarly, $\overline{X} \sim N(\mu_X, \sigma^2/m)$

Our hypotheses can be restated in terms of the difference $\mu_X - \mu_Y$:

$$H_0: \mu_X - \mu_Y = 0 \quad H_a: \mu_X - \mu_Y \neq 0$$

Thus we consider the difference $\bar{x} - \bar{y}$ as an estimate of $\mu_X - \mu_Y$. In the language of random variables, $\overline{X} - \overline{Y}$ is an *estimator* of $\mu_X - \mu_Y$.

Since our samples from X and Y are *independent*, the random variables $\overline{X}$ and $\overline{Y}$ are also independent. Mathematical theory tells us that:

1. The sum of *independent normal* random variables is normal (so we know that $\overline{X} - \overline{Y}$ is normal).
2. The mean (expected value) of the sum of random variables is the sum of the means of the terms (so we know that the mean of $\overline{X} - \overline{Y}$ is $\mu_X - \mu_Y$).
3. The variance of the sum or difference of *independent* random variables is the sum of the variances of the terms (so we know that $\text{Var}(\overline{X}) + \text{Var}(\overline{Y}) = \sigma^2/m + \sigma^2/n$).

Thus we have:

$$\overline{X} - \overline{Y} \sim N(\mu_X - \mu_Y, \ \sigma^2/m + \sigma^2/n).$$

Therefore

$$\frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma^2}{m} + \frac{\sigma^2}{n}}} \sim N(0,1) \quad \text{(i.e., is standard normal).}$$

If we knew σ^2, this would give us a test statistic to do inference on $\mu_X - \mu_Y$. But we don't know σ^2. We do, however, have two estimates of σ^2: the two *sample variances*

$$s_X^2 = \sum_{i=1}^{m} \frac{(x_i - \bar{x})^2}{m-1} \qquad \text{and} \qquad s_Y^2 = \sum_{i=1}^{n} \frac{(y_i - \bar{y})^2}{n-1}$$

Consistently with our notation $\bar{y}$ and $\bar{Y}$ above, we will use capital letters to refer to the underlying random variables (the *estimators* of σ^2):

$$S_X^2 = \sum_{i=1}^{m} \frac{(X_i - \bar{X})^2}{m-1} \qquad \text{and} \qquad S_Y^2 = \sum_{i=1}^{n} \frac{(Y_i - \bar{Y})^2}{n-1}$$

Which of these two estimators should we use? What seems better than picking one or the other is taking their average. But since they are come from possibly different sized samples, we use their *weighted* mean, yielding the *pooled estimator*

$$S^2 = \frac{(m-1)S_x^2 + (n-1)S_Y^2}{(m-1)+(n-1)} = \frac{1}{m+n-2}\left[\sum_{i=1}^{m}(x_i - \bar{x})^2 + \sum_{i=1}^{n}(y_i - \bar{y})^2\right].$$

So we consider the random variable $T = \dfrac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{\dfrac{S^2}{m} + \dfrac{S^2}{n}}}$

Mathematical theory (using the model assumptions) tells us that T has a t-distribution with n + m -2 degrees of freedom. If H_0 is true, then T is just

$$\frac{\bar{X} - \bar{Y}}{\sqrt{\dfrac{S^2}{m} + \dfrac{S^2}{n}}} = \frac{\bar{X} - \bar{Y}}{S\sqrt{\dfrac{1}{m} + \dfrac{1}{n}}}.$$

Summarizing: If H_0 is true, then the random value $T = \dfrac{\bar{X} - \bar{Y}}{S\sqrt{\dfrac{1}{m} + \dfrac{1}{n}}}$ has a t-distribution with n + m - 2 degrees of freedom. We can use fact to perform our hypothesis test: Calculate the value t of T determined by our sample. Calculate the corresponding p-value:

p = the probability of obtaining a value of T (having a t-distribution with n + m + 2 degrees of freedom) with absolute value greater than or equal to |t|.

If p is sufficiently small, we choose to reject the null hypothesis in favor of the alternate. Otherwise, we do not reject H_0 -- the evidence is consistent with it.

Example: Continuing with the example of comparing the two packages for teaching a new programming language, if we obtain sample mean 72.5 and sample standard deviation 10.3 for the first method, and sample mean and standard deviation 70.1 and 11.8, respectively, for the second method, then the *pooled sample variance* is

$$s^2 = [29(10.3^2) + 29(11.8^2)]/58 = 122.665,$$

so the *pooled standard deviation* is

$$s = \sqrt{(122.665)} = 11.075,$$

the *pooled standard error* (which is our estimate of the standard error of the random variable $\bar{X} - \bar{Y}$) is

$$\text{se}(\bar{x} - \bar{y}) = s\sqrt{\frac{1}{30} + \frac{1}{30}} = 2.86$$

and the t-statistic is

$$\frac{72.5 - 70.1}{11.075\sqrt{\frac{1}{30} + \frac{1}{30}}} = 2.4/2.86 = .8392$$

The corresponding p-value (two-tailed, using a t-distribution with 58 degrees of freedom) is 0.404825. This does not give us any evidence against the null hypothesis, so we have not detected any significant difference between the two packages -- we have no reason, based on the test scores, to choose one over the other.

We could also use the t-statistic to calculate a confidence interval for the difference $\mu_X - \mu_Y$ in the sample means. Suppose we want a 90% confidence interval. For 58 degrees of freedom, 90% of all values lie between - 1.67155 and + 1.67155. So for 90% of all samples satisfying the model assumptions,

$$-1.67155 < T < 1.67155.$$

In other words,

$$-1.67155 < \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{se(\bar{X} - \bar{Y})} < 1.67155.$$

A bit of algebra manipulation shows that this is equivalent to

$$(\bar{X} - \bar{Y}) - 1.67155se(\bar{X} - \bar{Y}) < \mu_X - \mu_Y < (\bar{X} - \bar{Y}) + 1.67155se(\bar{X} - \bar{Y}).$$

Evaluating this for our sample gives the endpoints

$$(72.5 - 70.1) \pm 1.67155(2.86)$$

for the confidence interval, resulting in confidence interval (-2.38, 7.18).

Note that we are *not* asserting that $\mu_X - \mu_Y$ lies in this interval. All we have done is use a procedure that, for 90% of all pairs of simple random samples of sizes n, chosen independently from the populations in question, will give an interval that does contain $\mu_X - \mu_Y$. Our sample could be one of the 10% yielding a confidence interval that does not contain $\mu_X - \mu_Y$.

Note also that the confidence interval contains zero. Thus our data are consistent with the possibility that $\mu_X - \mu_Y = 0$ -- in other words, that $\mu_X = \mu_Y$. (Note that this is the same conclusion we drew from the hypothesis test.)

References:

Ross, Sheldon M., Introduction to Probability and Statistics for Engineers and Scientists, Wiley, 1987

Wackerly, Dennis, William Mendenhall and Richard Scheafer, Mathematical Statistics with Applications, Duxbury, 1996