This is an introductory course in Analysis of Variance (ANOVA)

I. BRIEF OVERVIEW

What is Analysis of Variance?

Brief Answer: 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.

How is Analysis of Variance connected to Design of Experiments?

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
- The particulars of the context, such as:
  - The question of interest
  - Factors that may influence the variable of interest
  - Constraints such as time and budget.

II. MODEL ASSUMPTIONS FOR STATISTICAL PROCEDURES:

Assumptions about:

- the distributions of random variables involved, or
- how samples are chosen, or
- the type of relationship between the variables involved, or …

The essence of applying statistics: Selecting a model that does all of the following:

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

*G.E. Box:

All models are wrong; some are useful.

This means: 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.

III. REVIEW OF THE EQUAL VARIANCE TWO-SAMPLE T-TEST

(focusing on the model assumptions and why they are important.)

Note: 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, since it is the one that generalizes to basic Analysis of Variance methods

See handout "Review of Basic Statistical Concepts" if needed.

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.

- These are *population means*
- Do not confuse with *sample means* $\bar{x}$ and $\bar{y}$.

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, $n = m = 30$.)

Random variable X: "test score of an engineer from this company using the first package."

Random variable Y: "test score of an engineer from this company 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 plausible.

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):

(See references for more details.)

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

Notation: $X \sim N(\mu_X, \sigma^2)$ means: the random variable X is normally distributed with mean μ_X and variance σ^2.

Here:

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

$\overline{y}$ = sample mean of $y_1, y_2, \ldots y_n$ --- our best estimate of the mean μ_Y.

$\overline{y}$ is a value of the random variable $\overline{Y}$: "take a random sample of size n from Y and calculate its sample mean"

The distribution of the r.v. $\overline{Y}$ is called a *sampling distribution,* since the value of $\overline{Y}$ depends on the sample chosen.

Mathematical theory tells us: $\bar{Y}$ is normally distributed with mean μ_Y and variance σ^2/n:

$$\bar{Y} \sim N(\mu_Y, \sigma^2/n)$$

This conclusion uses the following facts (assumptions):

- $y_1, y_2, \ldots y_n$ *is a* <u>*random*</u> *sample*
- $Y \sim N(\mu_Y, \sigma^2)$.

Similarly (with the model assumptions),

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

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

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

The difference of sample means, $\bar{x} - \bar{y}$, is our best e*stimate* of $\mu_X - \mu_Y$.

In the language of random variables, $\bar{X} - \bar{Y}$ is an *estimator* of $\mu_X - \mu_Y$.

Since our samples from X and Y are *independent*, the random variables $\bar{X}$ and $\bar{Y}$ are also independent.

From mathematical theory:

1. The sum of *independent normal* random variables is normal (so we know that $\bar{X} - \bar{Y}$ 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 $\bar{X} - \bar{Y}$ 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 $\mathrm{Var}(\bar{X} - \bar{Y}) = \mathrm{Var}(\bar{X}) + \mathrm{Var}(\bar{Y}) = \sigma^2/m + \sigma^2/n$).

Thus:

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

Therefore

$$\frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma^2}{m} + \frac{\sigma^2}{n}}} \sim N(0,1)$$

(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*

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

These are values of the underlying random variables (the *estimators* of σ^2):

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

Which of these two estimators to use?

Better than either: Take their average -- their *weighted* mean (to take into account different sample sizes) to get 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

$$T = \frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{S^2}{m} + \frac{S^2}{n}}}$$

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

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

Summarize: If H_0 is true, then the random value

$$T = \frac{\overline{X} - \overline{Y}}{S\sqrt{\frac{1}{m} + \frac{1}{n}}}$$

has a t-distribution with n + m - 2 degrees of freedom.

Use this 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.

→ Note that this is *not* the same as saying H_a is true, and is also *not* the same as saying that H_0 is false – it just says that H_a appears to be the better option, given the evidence at hand.

Otherwise, we do not reject H_0 – we decide that the evidence does not contradict H_0 (i.e., is consistent with H_0.

→ Note that this is *not* the same as saying that H_0 is true, and it is also *not* the same as saying H_a is false – it's just saying that there is no reason to prefer H_a to H_0 , given the evidence at hand.

Example: (Comparing the two packages for teaching a new programming language, continued): 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

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

- 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

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

- The *t-statistic* is

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

- The *p-value* (two-tailed, using a t-distribution with 58 degrees of freedom) is 0.404825.

Conclusions:

- This does not give us any evidence against the null hypothesis
- So we have not detected any significant difference between the two packages.
- In other words, we have no reason, based just 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 a t-distribution with 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, for 90% of all samples satisfying the model assumptions

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

With some algebra:

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

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

(Remember: This is true for 90% of all samples satisfying the model assumptions -- possibly not for ours.)

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).

Please note: 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.)