Power Calculation for F-test

• If the population means $\mu_1, \mu_2, \ldots, \mu_r$ differ, we would like our F-test to reject H_0 with high probability.

Power =

• For any fixed α, the power of a test _____ as the sample size increases.

• For any fixed α, the power of the F-test is _____ when the $\mu_1, \mu_2, \ldots, \mu_r$ are more spread out.

• The distribution of F* under H_a is a _____

• In experiments, we often wish to determine the minimum sample size needed to have a specified power to detect certain differences among the population means.

Example: It may be important to reject H_0 when the largest and smallest μ_i differ by at least

Example: *Y* = Time until headache relief (in minutes), *r* = 4 different drugs

• Suppose the company wishes to detect when the mean times for the 4 drugs differ by at least 10 minutes with probability 0.80, using $\alpha = 0.05$.

• We also must specify (or at least guess) the value of σ. Suppose

Then

Table B.12 (pg. 1342-1343) tells us we need

• The R function `power.anova.test` does similar calculations, but <u>not</u> in terms of Δ.

<u>Investigating Differences Among Treatment Means</u>

• If the F-test is <u>not</u> significant, we conclude the true mean response is (or may be) the same for all factor levels and no further investigation is needed.

• If we conclude a <u>difference</u> in treatment means, we investigate further.

<u>Plots</u>: We can plot the sample means for each level using a bar graph (good with a qualitative factor) or a main effects plot (good with a quantitative factor).
• See SAS and R examples.

Inference About the Population Treatment Means

• Consider a CI for μ_i, the true mean response at level *i*:

Point estimate of μ_i is and

Using MSE to estimate σ^2,

→ 100(1 – α)% CI for μ_i is:

• SAS gives these using the CL option to the LSMEANS statement.

<u>Example</u>: 95% CI for the mean sales for package design 1 is:

Comparing Two Population Treatment Means

Estimating the Difference between μ_i and $\mu_{i'}$:

Point estimate is

So a 100(1 – α)% CI for $\mu_i - \mu_{i'}$ is:

• We also may test whether two particular treatment means are equal:

• This is done with a

• If we use the PDIFF option to the LSMEANS statement, SAS gives CIs for the difference between <u>each pair</u> of treatment means and the p-value for <u>each test</u> comparing two treatment means.

Note: These results are only valid if we are doing inference about one particular pair of treatment means. If we are doing many simultaneous comparisons, we must use different techniques.

SAS example: 95% CI for the difference between mean sales for package design 3 and mean sales for package design 4:

Contrasts

• A contrast is a linear combination of factor level means in which the coefficients sum to zero.
• Often useful for comparing several treatment means.

Previous example:

• In Kenton Foods data, package designs 1 and 3 used cartoons; designs 2 and 4 did not. To compare mean sales for the "cartoon designs" vs. the "non-cartoon designs", we could use the contrast:

• An unbiased estimator of a contrast *L* is

→ Using t procedures we can obtain a $100(1 - \alpha)\%$ CI for L or perform a t-test of

• SAS will do this with an ESTIMATE statement.

<u>Previous example</u>: 95% CI for difference in mean sales for the cartoon designs and mean sales of non-cartoon designs is:

Interpretation:

• Testing whether cartoon designs and non-cartoon designs have significantly different mean sales:

• Suppose we had wanted to test whether the non-cartoon designs had produced significantly <u>better</u> mean sales than the cartoon designs:

Simultaneous Inference

• Suppose we make 3 simultaneous comparisons, say:

• If we do three separate t-tests, each with $\alpha = 0.05$, then:

If H_0 is true in each case: The probability that we reject H_0 and incorrectly conclude H_a in at least one test is:

• So the family of tests has significance level

• Another concern: We must avoid "data snooping" – testing only those comparisons that are suggested by the data. This can wildly inflate the actual significance level!

• If we look at the data initially and then use that "early look" to decide which tests to perform, we are implicitly making comparisons.
• Solution: Use methods designed to make many inferences/comparisons simultaneously.
• These procedures are designed so that the family significance level is α

→

• The family significance level is also known as the

• This is as opposed to the
which is

Tukey's Multiple Comparison Procedure

• This compares all possible pairs of treatment means simultaneously:

• Tukey simultaneous CIs for <u>all</u> differences $\mu_i - \mu_{i'}$:

where

• We have confidence $(1 - \alpha)100\%$ that the <u>entire set</u> of these CIs contain the true pairwise differences between treatment means that they purport to contain.

Simultaneous Testing

We can test

• The Tukey procedure declares μ_i and $\mu_{i'}$ significantly different if

• Results of the Tukey procedure may be found in SAS.

Example (Kenton data): SAS output reveals CIs for all pairwise treatment mean differences, with family confidence coefficient 90%:

Testing: We can compare all pairs of treatment means, with family significance level $\alpha = 0.10$. SAS output shows:

Other Multiple Comparison Procedures

• The Scheffé procedure is designed for CIs for (and tests about) all possible contrasts.
• The family confidence level (significance level) is at least $1 - \alpha$ (at most α).
• Note that this class of inferences includes all possible pairwise comparisons, plus many more inferences.
• If only pairwise comparisons are needed, then Tukey's procedure is preferable.

• The Bonferroni procedure is good for testing about a few contrasts (must be specified before looking at the data or results – no data snooping allowed!)

• For each of g contrasts, L_1, …, L_g, the Bonferroni intervals are

For testing

against the two-sided alternatives, for each H_0, we reject if

where t^* is the estimated contrast divided by its standard error.

• The Bonferroni method is based on the Bonferroni inequality. Let A_i indicate that the i-th CI does not contain its parameter (or the i-th test has a Type I error):

• When testing about several contrasts, it is most useful to test <u>orthogonal</u> contrasts. Two contrasts

• Orthogonality implies that one contrast conveys no information about the other (i.e., no "overlapping information).
<u>Example</u>:

• When doing all pairwise comparisons of treatment means, the Tukey method is more efficient than Bonferroni. If only a few comparisons are of interest, the Bonferroni method may be better.
• When all (or very many) contrasts are of interest, the Scheffé procedure is best. When a few contrasts are of interest, Bonferroni may be better.
• Data snooping should not be used with the Bonferroni procedure, but data snooping is OK when Tukey or Scheffé is used, because these procedures intrinsically involve all possible inferences of a certain type anyway.

Other procedures: Dunnett's procedure is designed to compare several treatments to a "control" group simultaneously.
Example (Drug study):

Hsu's procedure: Selects "best" treatment and compares all others to the "best".

Checking Model Assumptions in ANOVA

• This is again done through an analysis of the

• We may check for outliers by examining the (internally) studentized residuals, which for the ANOVA model are:

Rule of Thumb:

• A formal t-test based on the externally studentized residuals is also available; see pg. 780 for details.

Graphical Tools: (1) Plots of Residuals vs. Fitted Values
Can check:

(2) Normal Q-Q plots of residuals
• Could do separate Q-Q plots for each factor level, if the sample sizes are quite large.
• Or could do a single Q-Q plot of all residuals (as long as error variances are judged roughly equal).
• Shapiro-Wilk test on the residuals can formally test for non-normality.

(3) If the data are gathered over time, we may plot the residuals against a time index (separately for each factor level) to check for dependence of errors across time.

• The equal variance assumption can be tested formally with the Brown-Forsythe test in SAS or R:

• The B-F test is based on doing the familiar F-test on the

• If the F-statistic is large and H_0 is rejected, then the variances are deemed unequal across populations.

SAS example (Kenton Foods data):

Common Remedies for Unequal Error Variances and/or Non-normal Errors

(1) Use weighted least squares and fit the ANOVA model with the regression approach.

(2) Transform the Response Variable:
Examples:

(3) Use a nonparametric alternative to the F-test called the Kruskal-Wallis test, which is based on the ranks of the data.