CHOOSING SAMPLE SIZES (See Section 3.6)

Several things need to be balanced in deciding how many observations to take on each treatment in an experiment. Primary among them are the cost of each additional observation, the budget available, the objectives of the experiment, and the number of observations needed to detect differences of interest. We will focus on the latter; in a real experiment, the results of these calculations may necessitate either reducing the objectives of the experiment or increasing the budget.

The method given here will be based on the idea of *power* of a statistical test: In general terms, the power of a test is the probability of rejecting H_o when a certain condition on the test statistic holds. For example, for a one-sample t-test for the mean of a population, with null hypothesis H_o: $\mu = 100$, you might be interested in the probability of rejecting H_o when $\mu = 105$, or when $|\mu - 100| > 5$, etc. In most real-life situations, there are reasonable such conditions. For example, if you can only measure the response to within 0.1 units, it doesn't really make sense to worry about false rejection when the actual value is within less than that amount of the value in the null hypothesis. Or some differences may be of no practical import -- for example, a medical treatment that extends life by 10 minutes is probably not worth it, but if it extends life by a year, it might be.

For an Analysis of Variance test, the *power of the test at Δ*, denoted $\pi(\Delta)$, is the probability of rejecting H_o when at least two of the treatments differ by Δ. The power $\pi(\Delta)$ depends on the sample size, the number of treatments v, the significance level α of the test, and the (population) error variance σ^2. The experimenter sets the values of Δ, $\pi(\Delta)$, v, and α, but in order to calculate sample size, an estimate of σ^2 is needed. This needs to be determined from a pilot experiment. However, we will see that if our estimate is lower than the actual variance, the power will be lower than $\pi(\Delta)$, so it is wise to estimate on the high side. On the other hand, if the estimate is too high, then the power will be higher than needed, and we will be rejecting H_o with high probability in cases where the difference is less than we really care about. The upper limit of a confidence interval for σ^2 thus seems like a reasonable choice of estimate. So we will start by looking at confidence intervals for σ^2.

Confidence Intervals for σ^2 (See Section 3.4.6)

It can be shown that $SSE/\sigma^2 \sim \chi^2(n\text{-}v)$. This allows us to adapt the usual confidence interval procedure to obtain one for σ^2: If we want, say, an upper 95% CI for σ^2, then we

find the 5th percentile $c_{0.05}$ of the $\chi^2(n-v)$ distribution, so that $P(SSE/\sigma^2 \geq c_{0.95}) = 0.95$. (Draw a picture!)

Equivalently: $P(SSE/c_{0.05} \geq \sigma^2) = 0.95$.

Replacing SSE by ssE, we get a 95% upper confidence interval $(0, SSE/c_{0.95})$ for σ^2, or a *95 upper confidence limit* $SSE/c_{0.05}$ for σ^2.

Example: In the battery experiment, the Minitab output gave ssE = 28413. Here, n = 16 and v = 4, so n - v = 12. Now $c_{0.05} = 5.226$ (5th percentile for $\chi^2(12)$), so the 95% upper confidence bound is 28413/5.226 = 5436.85. (Recall that the estimate for σ^2 was msE = 2368.)

"5436.85 is a 95% upper confidence bound for σ^2" means:

Calculating sample sizes for a given power

We will assume we are seeking equal sample sizes for each treatment, so will let r be the common value of the treatment sample sizes (i.e., all $r_i = r$). If we have specified a significance level α, then we will reject

$H_0: \tau_1 = \tau_2 = \ldots = \tau_v$

in favor of

H_a: At least two of the τ_i's differ

when

$msT/msE > F(v-1, n-v; \alpha)$.

Thus the power of the test is

$\pi(\Delta) = P(MST/MSE > F(v-1, n-v; \alpha))$

Recall that the test statistic MST/MSE has an F(v-1, n-v) distribution if H_0 is true. If H_0 is not true, then the distribution of MST/MSE has what is called a *noncentral F-distribution*. (This is the distribution of the quotient of two "non-central chi-squared" random variables.) This distribution is defined in terms of a *noncentrality parameter* δ^2, which under our assumptions is given by

$$\delta^2 = \frac{r\sum_{i=1}^{v}(\tau_i - \bar{\tau}.)^2}{\sigma^2}$$

It turns out that the hardest situation to detect (therefore the "limiting case" on which the sample size calculations depend) is when the effects of two of the factor levels differ by Δ and the rest are all equal and midway between these two.

Exercise: In this limiting case, the formula for δ^2 reduces to

$$\delta^2 = \frac{r\Delta^2}{2\sigma^2}.$$

Thus

$$r = \frac{2\sigma^2\delta^2}{\Delta^2}.$$

Power for the noncentral F distribution is given in tables as a function of $\phi = \delta/\sqrt{\nu}$. So in terms of ϕ,

$$r = \frac{2\nu\sigma^2\phi^2}{\Delta^2}.$$

However, the tables need to be used iteratively, since the denominator degrees of freedom n - v = v(r-1) depend on r. A procedure for doing this is given on p. 52 of the text.