Lecture 7
Topic 5: Multiple Comparisons (means separation)

ANOVA: H_0: $\mu_1 = \mu_2 = ... = \mu_t$
H_1: The mean of at least one treatment group is different

If there are more than two treatments in the experiment, further analysis is required to determine *which* means are significantly different. There are two strategies:

1. **Planned, single d.f. F tests (orthogonal contrasts)**

   Motivated by the treatment structure
   Independent
   Powerful and precise
   Limited to (t – 1) comparisons
   Well-defined method

2. **Multiple comparisons (means separation)**

   Motivated by the data
   Useful when no particular relationship exists among treatments
   Up to unlimited comparisons
   Many, many methods/philosophies to choose from

Error rates

Selection of the most appropriate multiple comparison test is heavily influenced by the ***error rate***. Recall that a Type I error occurs when one incorrectly rejects a true null hypothesis H_0.

The **Type I error rate** is the fraction of times a Type I error is made.

In a single comparison, this is α. When comparing three or more treatment means:

1. **Comparison-wise Type I error rate (CER)**

   The number of Type I errors, divided by the total number of comparisons.

2. **Experiment-wise Type I error rate (EER)**

   The number of experiments in which **at least** one Type I error occurs, divided by the total number of experiments.

Example: An experimenter conducts an experiment with 5 treatments.

In such an experiment, there are 10 possible pairwise comparisons that can be made:

$$\text{Total possible pairwise comparisons:} \quad \mathrm{p} = \frac{t(t-1)}{2}$$

Suppose that there are no true differences among the treatments (i.e. H_0 is true), but that one Type I error is made.

CER = (1 Type I error) / (10 comparisons) = 0.1 or 10%

EER = (1 experiments with a Type I error) / (1 experiment) = 1 or 100%

Things to consider:

1. EER is the probability of there being a Type I error somewhere in the experiment. As the number of treatments increases, the EER → 100%.

2. To maintain a low EER, the CER has to be kept very low. Conversely, a reasonable CER will inflate the EER to a potentially unacceptable level.

3. The relative importance of controlling these two Type I error rates depends on the objectives of the study:

   When incorrectly rejecting one comparison jeopardizes the entire experiment or when the consequence of incorrectly rejecting one comparison is as serious as incorrectly rejecting a number of comparisons, EER control is more important.

   When one erroneous conclusion will not affect other inferences in an experiment, CER control is more important.

4. Different multiple comparison procedures have been developed based on different philosophies regarding control of these two kinds of error.

Computing EER

So you set CER = α … what is EER?

The EER is difficult to compute because, for a given set of data, Type I errors are not independent. But it *is* possible to compute an upper bound for the EER by assuming that the probability of a Type I error for any single comparison is α and is independent of all other comparisons. In that case:

$$\boxed{\text{Upper bound EER} = 1 - (1 - \alpha)^p \text{ where } p = \frac{t(t-1)}{2}\text{, as before}}$$

Example: For 10 treatments and $\alpha = 0.05$:

$$p = \frac{t(t-1)}{2} = \frac{10(10-1)}{2} = 45$$

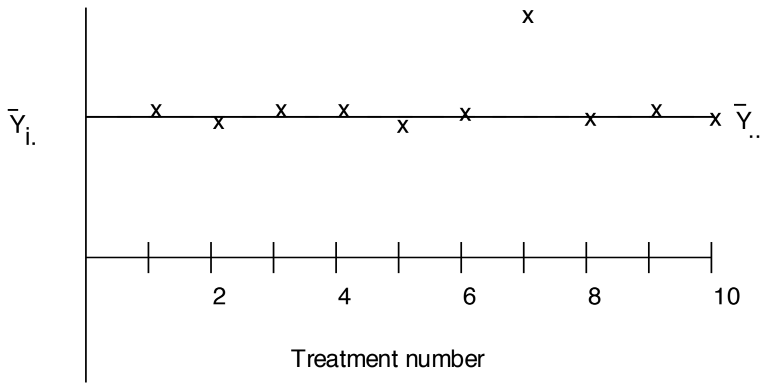
$$\text{Upper bound EER} = 1 - (1 - 0.05)^{45} = 0.90$$

This formula may also be used to determine a value for α for some fixed maximum EER.

$$0.1 = 1 - (1 - \alpha)^{45}$$
$$(1 - \alpha)^{45} = 0.9$$
$$(1 - \alpha) = 0.9^{(1/45)}$$
$$\alpha = 0.002$$

Partial null hypothesis

Suppose there are 10 treatments, one of which shows a significant effect while the other 9 are approximately equal:



ANOVA will probably reject H_0.

Even though one mean is truly different, there is still a chance of making a Type I error in *each pairwise comparison* among the 9 similar treatments.

An upper bound the EER is computed by setting t = 9 in the above formula:

$$p = \frac{t(t-1)}{2} = \frac{9(9-1)}{2} = 36$$

$$\text{Upper bound EER} = 1 - (1 - 0.05)^{36} = 0.84$$

Interpretation: The experimenter will incorrectly conclude that two truly similar effects are different **84% of the time**. This is called the experiment-wise error rate under a partial null hypothesis.

Some terminology:

CER = comparison-wise error rate
EERC = experiment-wise error rate under a complete null hypothesis (standard EER)
EERP = experiment-wise error rate under a partial null hypothesis
MEER = maximum experiment-wise error rate under any complete or partial null hypothesis.

Multiple comparisons tests

Statistical methods for making two or more inferences while controlling cumulative Type I error rates are called ***simultaneous inference methods***:

1. Fixed-range tests: Those which provide confidence intervals and tests of hypotheses
2. Multiple-range tests: Those which provide only tests of hypotheses

Equal replications. Results (mg shoot dry weight) of an experiment (CRD) to determine the effect of seed treatment by different acids on the early growth of rice seedlings.

Treatment	Replications					Mean
Control	4.23	4.38	4.10	3.99	4.25	4.19
HCl	3.85	3.78	3.91	3.94	3.86	3.87
Propionic	3.75	3.65	3.82	3.69	3.73	3.73
Butyric	3.66	3.67	3.62	3.54	3.71	3.64

t = 4, r = 5, overall mean = 3.86

Source	df	SS	MS	F
Total	19	1.0113		
Treatment	3	0.8738	0.2912	33.87
Error	16	0.1376	**0.0086**	

Unequal replications. Results (lbs/animal·day) of an experiment (CRD) to determine the effect of different forage genotypes on animal weight gain.

Treatment	Replications (Animals)								r	Mean
Control	1.21	1.19	1.17	1.23	1.29	1.14			6	1.205
Forage-A	1.34	1.41	1.38	1.29	1.36	1.42	1.37	1.32	8	1.361
Forage-B	1.45	1.45	1.51	1.39	1.44				5	1.448
Forage-C	1.31	1.32	1.28	1.35	1.41	1.27	1.37		7	1.330

t = 4, r = variable, overall mean = 1.336

Source	df	SS	MS	F
Total	25	0.2202		
Treatment	3	0.1709	0.05696	25.41
Error	22	0.0493	**0.00224**	

Fixed-range tests

These tests provide a **single range** for testing all differences in balanced designs and can provide confidence intervals.

LSD → Dunnett → Tukey → Scheffe

Less conservative → More conservative
More likely to declare differences → Less likely to declare differences
Higher Type I error rates → Lower Type I error rates
Higher power → Lower power

Least significant difference (LSD), the repeated t test

One of the oldest, simplest, and most widely **misused** multiple pairwise comparison tests.

The LSD test declares the difference between means $\overline{Y}_i$ and $\overline{Y}_j$ of treatments T_i and T_j to be significant when:

$$|\overline{Y}_i - \overline{Y}_j| > \text{LSD, where}$$

$$LSD = t_{\frac{\alpha}{2},df_{MSE}} \sqrt{MSE\left(\frac{1}{r_1} + \frac{1}{r_2}\right)} \quad \text{for unequal r}$$

$$LSD = t_{\frac{\alpha}{2},df_{MSE}} \sqrt{MSE\frac{2}{r}} \quad \text{for equal r}$$

Seed treatment data: MSE = 0.0086 and df_{MSE} = 16.

$$LSD = t_{\frac{\alpha}{2},df_{MSE}} \sqrt{MSE\frac{2}{r}} = 2.120\sqrt{0.0086\frac{2}{5}} = 0.1243$$

So, if $|\overline{Y}_i - \overline{Y}_j| > 0.1243$, they are declared significantly different.

Control	4.19
HCl	3.87
Propionic	3.73
Butyric	3.64

Treatment	Mean	LSD
Control	4.19	a
HCl	3.87	b
Propionic	3.73	c
Butyric	3.64	c

All acids reduced shoot growth.
The reduction was more severe with butyric and propionic acid than with HC1.
We do not have evidence to conclude that propionic acid is different in its effect than butyric acid.

When treatments are equally replicated, only one LSD value
is required to test all possible comparisons.

Forage data: MSE = 0.00224 and df_{MSE} = 22.

In cases of unequal replication, different LSD values must be calculated for each comparison involving different numbers of replications.

The 5% LSD for comparing the control with Feed B:

$$LSD = t_{\frac{\alpha}{2},df_{MSE}}\sqrt{MSE\left(\frac{1}{r_{Cont}}+\frac{1}{r_B}\right)} = 2.074\sqrt{0.00224\left(\frac{1}{6}+\frac{1}{5}\right)} = 0.0594$$

A vs. Control	= 0.0531
A vs. B	= 0.0560
A vs. C	= 0.0509
B vs. C	= 0.0575
C vs. Control	= 0.0546

Treatment	Mean	LSD
Feed B	1.45	a
Feed A	1.36	b
Feed C	1.33	b
Control	1.20	c

At the 5% level, we conclude all feeds cause significantly greater weight gain than the control. Feed B causes the highest weight gain; Feeds A and C are equally effective.

Confidence intervals

The $(1 - \alpha)$ confidence limits of the quantity $(\mu_A - \mu_B)$ are given by:

$$(1 - \alpha) \text{ CI for } (\mu_A - \mu_B) = (\overline{Y}_A - \overline{Y}_B) \pm \text{LSD}$$

General considerations for LSD

The LSD test is much safer when the means to be compared are selected *in advance* of the experiment (i.e. before looking at the data).

The LSD test is the only test for which CER equals α. This is often regarded as too liberal.

It has been suggested that the EEER can be maintained at α by performing the overall ANOVA test at the α level and making further comparisons *if and only if* the F test is significant (**Fisher's Protected LSD test**). However, it was then demonstrated that this assertion is false if there are more than three means:

A preliminary F test controls only the EERC, not the EERP.

Bonferroni to the rescue...

Again consider the case of 5 treatments and thus 5*4/2 = 10 pairwise comparisons (i.e. hypotheses):

$$\alpha = 0.05$$
$$\alpha_{\text{Bon}} = 0.05/10 = 0.005$$

$$\text{Upper bound EER} = 1 - (1 - 0.005)^{10} = 0.0488$$

Dunnett's Test

Pairwise comparison of a control to all other treatment means,
while holding MEER ≤ α.

This test uses the t* statistic, a modified t statistic based on the number of comparisons to be made (p = number of treatment means, excluding the control).

$$DLSD = t^*_{\frac{\alpha}{2},p,df_{MSE}} \sqrt{MSE\left(\frac{1}{r_1}+\frac{1}{r_2}\right)} \quad \text{for unequal r } (r_0 \neq r_i)$$

$$DLSD = t^*_{\frac{\alpha}{2},p,df_{MSE}} \sqrt{MSE\frac{2}{r}} \quad \text{for equal r } (r_0 = r_i)$$

Seed treatment data: MSE = 0.0086, df_{MSE} = 16, and p = 3.

$$DLSD = t^*_{\frac{\alpha}{2},p,df_{MSE}} \sqrt{MSE\frac{2}{r}} = 2.59\sqrt{0.0086\frac{2}{5}} = 0.1519$$

(**DLSD = 0.1519 > LSD = 0.1243**)

The smallest difference between the control and any acid treatment is:

Control - HC1 = 4.19 – 3.87 = 0.32 > 0.1519

All other differences, being larger, are also significant.

The 95% simultaneous confidence intervals for all three differences take the form:

$(1 - \alpha)$ CI for $(\mu_0 - \mu_i) = (\overline{Y}_0 - \overline{Y}_i) \pm$ DLSD

Control – Butyric	= 0.32 ± 0.15
Control – HC1	= 0.46 ± 0.15
Control – Propionic	= 0.55 ± 0.15

We have 95% confidence that the 3 true differences fall **simultaneously** within the above ranges.

Animal forage data: MSE = 0.00224, df$_{\text{MSE}}$ = 22, and p = 3.

When treatments are not equally replicated, there are different DLSD values for each of the comparisons.

The 5% DSLS to compare the control with Feed-C:

$$DLSD = t^*_{\frac{\alpha}{2},p,df_{MSE}} \sqrt{MSE\left(\frac{1}{r_0}+\frac{1}{r_1}\right)} = 2.517\sqrt{0.00224\left(\frac{1}{6}+\frac{1}{7}\right)} = 0.0663$$

Since $|\overline{Y}_0 - \overline{Y}_C| = 0.125 > 0.06627$, the difference is significant. All other differences with the control, being larger than this, are also significant.

Tukey's *w* procedure

All possible pairwise comparisons, while holding MEER ≤ α.

Sometimes called the "honestly significant difference" (HSD) test, Tukey's controls the MEER *when the sample sizes are equal*. Instead of t or t*, it uses the statistic $q_{\alpha,p,df_{MSE}}$:

$$q_{\alpha,p,df_{MSE}} = \frac{\overline{Y}_{MAX} - \overline{Y}_{MIN}}{s_{\overline{Y}}}$$

The critical difference in this method is labeled *w*:

$$w = q_{\alpha,p,df_{MSE}} \sqrt{\frac{MSE}{2}\left(\frac{1}{r_1} + \frac{1}{r_2}\right)} \quad \text{for unequal r}$$

$$w = q_{\alpha,p,df_{MSE}} \sqrt{\frac{MSE}{r}} \quad \text{for equal r}$$

We do not multiply MSE by a factor of 2 *because Table A-8 (class website) already includes the factor √2 in its values*:

For p = 2, df = ∞, and α= 5%, the critical value is 2.77 = 1.96 * √2

Tukey critical values are larger than those of Dunnett because the Tukey family of contrasts is larger (all pairs of means).

Seed treatment data: MSE = 0.0086, df_{MSE} = 16, and p = 4.

$$w = q_{\alpha,p,df_{MSE}} \sqrt{\frac{MSE}{r}} = 4.05\sqrt{\frac{0.0086}{5}} = 0.1680$$

(**w = 0.1680 > DLSD = 0.1519 > LSD = 0.1243**)

Treatment	Mean	*w*	
Control	4.19	a	
HCl	3.87	b	
Propionic	3.73	**b**	**c**
Butyric	3.64		c

Animal forage data: MSE = 0.00224, df_{MSE} = 22, and p = 4.

The 5% *w* for the contrast between the Control and Feed-C:

$$w = q_{\alpha,p,df_{MSE}}\sqrt{\frac{MSE}{2}\left(\frac{1}{r_{Cont}}+\frac{1}{r_C}\right)} = 3.93\sqrt{\frac{0.00224}{2}\left(\frac{1}{6}+\frac{1}{7}\right)} = 0.0732$$

Since $|\overline{Y}_{Cont} - \overline{Y}_C|$ = 0.125 > 0.0731, it is significant. As in the LSD, the only pairwise comparison that is not significant is that between Feed C ($Y_C = 1.330$) and Feed A ($Y_A = 1.361$).

Scheffe's F test

Compatible with the overall ANOVA F test: Scheffe's never declares a contrast significant if the overall F test is nonsignificant.

Scheffe's test controls the MEER for **ANY** set of contrasts. This includes *all possible pairwise and group comparisons*.

Since this procedure allows a larger number of comparisons, it is less sensitive than other multiple comparison procedures.

For pairwise comparisons, the Scheffe critical difference (SCD) has a similar structure as that described for previous tests:

$$SCD = \sqrt{df_{Trt}F_{\alpha,df_{Trt},df_{MSE}}}\sqrt{MSE\left(\frac{1}{r_1}+\frac{1}{r_2}\right)} \quad \text{for unequal r}$$

$$SCD = \sqrt{df_{Trt}F_{\alpha,df_{Trt},df_{MSE}}}\sqrt{MSE\frac{2}{r}} \quad \text{for equal r}$$

Seed treatment data: MSE = 0.0086, df_{Trt} = 3, df_{MSE} = 16:

$$SCD = \sqrt{df_{Trt}F_{\alpha,df_{Trt},df_{MSE}}}\sqrt{MSE\frac{2}{r}} = \sqrt{3(3.24)0.0086\frac{2}{5}} = 0.1829$$

(SCD = 0.1829 > w = 0.1680 > DLSD = 0.1519 > LSD = 0.1243)

The table of means separations:

Treatment	Mean	F_s	
Control	4.19	a	
HCl	3.87	b	
Propionic	3.73	b	c
Butyric	3.64		c

Animal forage data: MSE = 0.00224, df_{Trt} = 3, df_{MSE} = 22.

The 5% SCD for the contrast between the Control and Feed-C:

$$SCD = \sqrt{df_{Trt} F_{\alpha, df_{Trt}, df_{MSE}}} \sqrt{MSE\left(\frac{1}{r_1} + \frac{1}{r_2}\right)} = \sqrt{3(3.05)0.00224\left(\frac{1}{6} + \frac{1}{7}\right)} = 0.0796$$

Since $|\overline{Y}_{Cont} - \overline{Y}_C| = 0.125 > 0.0796$, it is significant.

Scheffe's procedure is also readily used for interval estimation:

$$(1 - \alpha) \text{ CI for } (\mu_0 - \mu_i) = (\overline{Y}_0 - \overline{Y}_i) \pm \text{SCD}$$

The resulting intervals are **simultaneous** in that the probability is at least (1 – α) that all of them are true simultaneously.

Scheffe's F test for group comparisons

The most important use of Scheffe's test is for
arbitrary comparisons among groups of means.

To make comparisons among groups of means, you first define a contrast, as in Topic 4:

$$Q = \sum_{i=1}^{t} c_i \overline{Y}_i \text{, with the constraint that } \sum_{i=1}^{t} c_i = 0 \text{ (or } \sum_{i=1}^{t} r_i c_i = 0 \text{ for unequal r)}$$

We reject the null hypothesis (H_0) that the contrast Q = 0 if the absolute value of Q is larger than some critical value F_S:

$$\text{Critical value } F_S = \sqrt{df_{Trt} F_{\alpha, df_{Trt}, df_{MSE}}} \sqrt{MSE \sum_{i=1}^{t} \frac{c_i^2}{r_i}}$$

(The previous pairwise expressions are for the particular contrast 1 vs. -1.)

Example: If we wish to compare the control to the average of the three acid treatments, the contrast coefficients are (+3, –1, –1, –1). In this case:

$$Q = \sum_{i=1}^{t} c_i \overline{Y}_i = 4.190(3) + 3.868(-1) + 3.728(-1) + 3.640(-1) = 1.334$$

The critical F_s value for this contrast is:

$$F_S = \sqrt{df_{Trt} F_{\alpha, df_{Trt}, df_{MSE}}} \sqrt{MSE \sum_{i=1}^{t} \frac{c_i^2}{r_i}} = \sqrt{3(3.24)0.0086 \frac{3^2 + (-1)^2 + (-1)^2 + (-1)^2}{5}} = 0.4479$$

Since |Q| = 1.334 > 0.4479 = F_S, we reject H_0. The average of the control (4.190 mg) is significantly larger than the average of the three acid treatments (3.745 mg).

Multiple-stage tests (MSTs) / Multiple-range tests

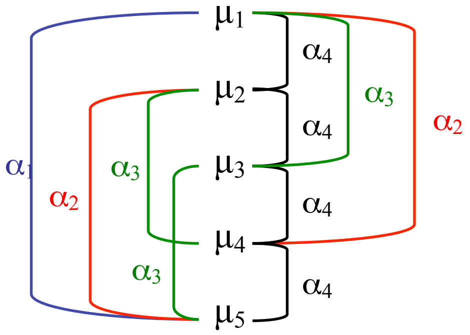
Allow simultaneous hypothesis tests of greater power by forfeiting the ability to construct simultaneous confidence intervals.

Duncan → Student-Newman-Keuls (SNK) → REGWQ

All three use the Studentized range statistic (q_{α}), and all three are result-guided.

> With means arranged in order, an MST provides critical distances or ranges that become smaller as the pairwise means to be compared become closer together in the array. Such a strategy allows the researcher to allocate test sensitivity where it is most needed, in discriminating neighboring means.

The general strategy:



$$\alpha_1 < \alpha_2 < \alpha_3 < \ldots < \alpha_{t-1}$$

"Confidence" is replaced by the concept of "protection levels"

So if a difference is detected at one level of the test, the researcher is justified in separating means at a finer resolution with less protection (i.e. with a higher α).

Duncan's multiple range test

As the test progresses, Duncan's method uses a variable significance level (α_{p-1}) depending on the number of means involved:

$$\alpha_{p-1} = 1 - (1 - \alpha)^{p-1}$$

Despite the level of protection offered at each stage, MEER is uncontrolled. The higher power of Duncan's method compared to Tukey's is due to its higher Type I error rate.

Duncan critical ranges (R_p):

$$R_p = q_{\alpha_{p-1},p,df_{MSE}} \sqrt{\frac{MSE}{r}}$$

For the seed treatment data:

p	**2**	**3**	**4**
$q_{\alpha_{p-1},p,16}$	3.00	3.15	3.23
R_p	**0.124**	0.131	0.134

Identical to LSD for adjacent means (LSD = 0.124).

Duncan's used to be the most popular method of means separation, but many journals no longer accept it. It is not recommended.

The Student-Newman-Keuls (SNK) test

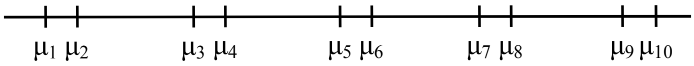
As the test progresses, SNK uses a fixed significance level (α), which is always less than or equal to Duncan's variable significance level:

$$\alpha_{SNK} = \alpha \leq 1 - (1 - \alpha)^{p-1}$$

More conservative than Duncan's, holding EERC ≤ α.
Accepted by some journals that reject Duncan's.

Poor behavior in terms of EERP and MEER.
Not recommended.

Assume the following partial null hypothesis:



The SNK method reduces to five independent tests, one for each pair, by LSD. The probability of at least one false rejection is:

$$1 - (1 - \alpha)^5 = 0.23$$

As the number of means increases, MEER → 1.

To find the SNK critical range (W_p) at each level of the analysis:

$$W_p = q_{\alpha,p,df_{MSE}}\sqrt{\frac{MSE}{r}}$$

For the seed treatment data:

p	**2**	**3**	**4**
$q_{0.05,p,16}$	3.00	3.65	4.05
R_p	**0.124**	0.151	0.168

Again, identical to LSD for adjacent means (LSD = 0.124).

The Ryan, Einot, Gabriel, and Welsh (REGWQ) method

Not as well known as the others, REGWQ method appears to be among the most powerful step-down multiple range tests and is recommended by some software packages (e.g. SAS) for *equal replication* (i.e. balanced designs).

Controls MEER by setting:

$$\alpha_{p-1} = 1 - (1 - \alpha)^{p/t} \text{ for } p < (t - 1) \quad \text{and} \quad \alpha_{p-1} = \alpha \text{ for } p \geq (t - 1)$$

Assuming the sample means have been arranged in descending order from $\overline{Y}_1$ to $\overline{Y}_t$, the homogeneity of means $\overline{Y}_i,\ldots,\overline{Y}_j$, with i < j, is rejected by REGWQ if:

$$| \overline{Y}_i - \overline{Y}_j | > q_{\alpha_{p-1},p,df_{MSE}} \sqrt{\frac{MSE}{r}}$$

For the seed treatment data:

p	2	3	4
α_{p-1}	0.025	0.05	0.05
$q_{0.05,p,16}$	3.49	3.65	4.05
R_p	**0.145**	0.151	**0.168**
	>SNK <Tukey	=SNK	=SNK =Tukey

Tukey *w* = 0.168

The difference between the HCl and Propionic treatments is declared significant with SNK but not with REGWQ (3.87 - 3.73 < 0.145).

Treatment	Mean	*REGWQ*	
Control	4.19	a	
HCl	3.87	b	
Propionic	3.73	b	c
Butyric	3.64		c

Some suggested rules of thumb:

1. When in doubt, use Tukey.

2. Use Dunnett's (more powerful than Tukey's) if you only wish to compare each treatment level to a control.

3. Use Scheffe's if you wish to "mine" your data.

One final point to note is that severely unbalanced designs can yield very strange results:

Treatment	Data											Mean		
A	10	17										13.5		NS
B	10	11	12	12	13	14	15	16	16	17	18	14.0	*	
C	14	15	16	16	17	18	19	20	20	21	22	18.0		
D	16	21										18.5		

Data from ST&D page 200.