Comparing Multiple Independent Populations - Summary

1. ANOVA

If we have simple random samples from more than two treatments (say, t treatments), and wish to test:

$H_0 : \mu_1 = \mu_2 = ... = \mu_t$
vs.
H_A: At least one mean differs from one other mean,

and we are willing to assume:

- The data are independent within and between treatments
- The variances are the same for all treatments
- Each treatment has a normal distribution,

then we can use ANOVA. The basic idea is to compare variability between treatments to the variability within treatments. Using the following notation:

Let t be the number of treatments.
Let i index the treatments.
Let n_i be the number of observations in treatment i.
Let $N = \sum_{i=1}^{t} n_i$ be the total sample size.
Let y_{ij} be observation j from treatment i.
Let $\bar{y}_{i.} = \frac{\sum_{j=1}^{n_i} y_{ij}}{n_i}$ be the sample mean for treatment i.
Let $\bar{y}_{..} = \frac{\sum_{i=1}^{t}\sum_{j=1}^{n_i} y_{ij}}{N}$ be the sample grand mean.

we can compute the sums of squares total, treatment, and error:

$SSTot = \sum_{i=1}^{t}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{..})^2$
$SSTrt = \sum_{i=1}^{t}\sum_{j=1}^{n_i}(\bar{y}_{i.} - \bar{y}_{..})^2$
$SSE = \sum_{i=1}^{t}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2$

and then summarize what we need in an ANOVA table:

Source	SS	df	MS	F	p-value
Treat (between)	SSTrt	$df_{Trt} = t - 1$	MSTrt = SSTrt/df_{Trt}	F = MSTrt/MSE	$p = P(F_{df_{Trt},df_E} > F)$
Error (within)	SSE	$df_E = N - t$	MSE = SSE/df_E		
Total	SSTot	$df_{Tot} = N - 1$			

P-values can be found using an F table. Important graphs include side-by-side dotplots of the original data, a residuals vs fitted values plot (to check equal variance), and a QQ plot of residuals (to check normality).

2. Multiple Comparisons

If we do not reject the null, we're done. But if we reject, we can use pairwise tests to compare individual means. The procedure is to compare treatments i and i' with either a special t-test:

$$t = \frac{\bar{y}_{i.} - \bar{y}_{i'.}}{\sqrt{MSE(1/n_i + 1/n_{i'})}}$$

which we would compare to a t_{df_E}. Or, to make a $100(1-\alpha)\%$ CI on the difference:

$$\bar{y}_{i.} - \bar{y}_{i'.} \pm t_{df_E, \alpha/2}\sqrt{MSE(1/n_i + 1/n_{i'})}$$

Often this information is summarized by sorting the treatment means from largest to smallest, and then adding letter codes. Two treaments share a letter if they do not differ significantly.