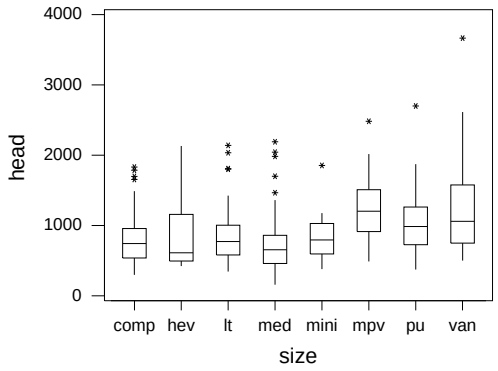
Comparing Several Means

Sometimes we wish to compare the means for groups of subjects from several populations.

Consider the crashtest data on head injury for different types of vehicle (the `size` variable):



There seem to be some differences between groups, but could this just be due to chance?

The Model for One-Way Analysis of Variance (ANOVA)

Suppose there are k groups, with n_j subjects in the jth group. We'll use $y_{i,j}$ to denote the value observed for the i'th subject in group j.

We model this data in terms of means for each group, μ_j, plus residuals for each data point:

$$y_{i,j} = \mu_j + \epsilon_{i,j}$$

We will assume that the residuals, $\epsilon_{i,j}$, are independent, and are normally-distributed with mean zero and standard deviation σ.

From the data, we will learn something about the mean parameters, μ_j, and about σ. We can test hypotheses about the means, or find confidence intervals.

The Null Hypothesis of Equal Means

The null hypothesis for one-way ANOVA is that the population means for all groups are equal:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$

$$H_a : \text{not all the } \mu_j \text{ are equal}$$

Note: we are looking here for *some* difference among the means, not for any difference in particular.

When we have only two groups, this is the same as the null hypothesis for the two sample t test. In this case, the ANOVA hypothesis test will be equivalent to a two-sided t test with pooled variance estimate.

Estimating the ANOVA Parameters

The one-way ANOVA model $y_{i,j} = \mu_j + \epsilon_{i,j}$ has parameters $\mu_1, \ldots, \mu_k$ and σ^2 (the variance of the ϵ_{ij}).

We can estimate these parameters as follows:

$$\mu_j \text{ by } \bar{y}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{i,j}, \qquad j = 1, \ldots, k$$

$$\sigma^2 \text{ by } s_p^2 = \frac{(n_1-1)s_1^2 + \cdots + (n_k-1)s_k^2}{(n_1-1) + \cdots + (n_k-1)}$$

$$= \frac{1}{N-k} \sum_{j=1}^{k} \sum_{i=1}^{n_j} (y_{i,j} - \bar{y}_j)^2$$

where $N = n_1 + \cdots + n_k$.

These estimators are unbiased.

Sums of Squares and Mean Squares

The name "Analysis of Variance" comes from looking at how how the overall sum of squared deviations from the mean can be decomposed into parts. This is related to how the variance of a sum of independent random variables is the sum of the variances of the terms.

The *total sum of squares* is

$$\text{SSTotal} = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (y_{i,j} - \bar{y})^2$$

where $\bar{y}$ is the overall sample mean (called $\bar{y}_G$ in the text).

This sum of squares is associated with a *total degrees of freedom* of $\text{DFTotal} = N - 1$.

The associated *mean square* is

$$\text{MSTotal} = \text{SSTotal} / \text{DFTotal}$$

which is the sample variance for the entire data set.

Decomposing the Sum of Squares

The total sum of squares can be decomposed into variation between groups and variation within groups:

$$\text{SSTotal} = \text{SST} + \text{SSE}$$

where SST is the *between group sum of squares*, due to different "treatments":

$$\text{SST} = \sum_{j=1}^{k} n_j (\bar{y}_j - \bar{y})^2$$

and SSE is the *sum of squares for the error*:

$$\text{SSE} = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (y_{i,j} - \bar{y}_j)^2$$

The associated degrees of freedom are $\text{DFT} = k - 1$ and $\text{DFE} = N - k$.

Note that $\text{MSE} = \text{SSE}/\text{DFE} = s_p^2$.

$R^2 = \text{SST}/\text{SSTotal}$ is the proportion of the total sum of squares explained by the variation between groups.

The F Statistic

To test the null hypothesis that $\mu_1 = \cdots = \mu_k$, we look at the statistic

$$F = \frac{\text{MST}}{\text{MSE}}$$

Under the null hypothesis, both $\text{MSE} = s_p^2$ and $\text{MST} = \text{SST}/\text{DFT}$ and have means of σ^2, but they vary randomly, giving a distribution for F that depends on the degrees of freedom of the numerator $(k-1)$ and denominator $(N-k)$.

This is called the F distribution. We compute F for our data and look in a table (or use minitab) to find the probability of getting a value this large or larger under the null hypothesis. This is our P-value.

When there are two groups, the result is the same as a two-sided two-sample t test with pooled variance estimate.

ANOVA for the Crashtest Data: Head Injury for Different Vehicle Types

```
One-way Analysis of Variance

Analysis of Variance for head
Source     DF        SS        MS        F        P
size        7  11226093   1603728     8.58    0.000
Error     332  62063350    186938
Total     339  73289443

                                  Individual 95% CIs For Mean
                                  Based on Pooled StDev
Level      N      Mean     StDev  ------+---------+---------+---------+
comp      83     784.9     339.6      (--*---)
hev       14     837.9     489.7  (---------*--------)
lt        73     826.2     375.6       (---*---)
med       58     740.3     427.8   (----*---)
mini      14     853.6     366.4   (--------*--------)
mpv       32    1245.1     443.6                     (-----*-----)
pu        36    1050.5     460.6               (-----*-----)
van       30    1243.6     685.6                     (-----*-----)
                                  ------+---------+---------+---------+
Pooled StDev =    432.4               750       1000      1250      1500
```

But are the assumptions behind the ANOVA model satisfied here? Recall that the units here are dummies, and usually there is a dummy in the driver's seat and one in the passenger's seat. How could this be fixed?