# Section 8.1: Analysis of Variance

Eric D. Nordmoe

Math 261
Biostatistics
Kalamazoo College

# Outline

- Overview of Analysis of Variance
- The Basic ANOVA
- The ANOVA Model
- Checking Conditions

# The Basic Idea of ANOVA

When we ask if a set of sample means gives evidence for differences among the population means, what matters is not how far apart the sample means are but how far apart they are relative to the variability of individual observations.

*Baldi & Moore, 3rd ed., p.606*

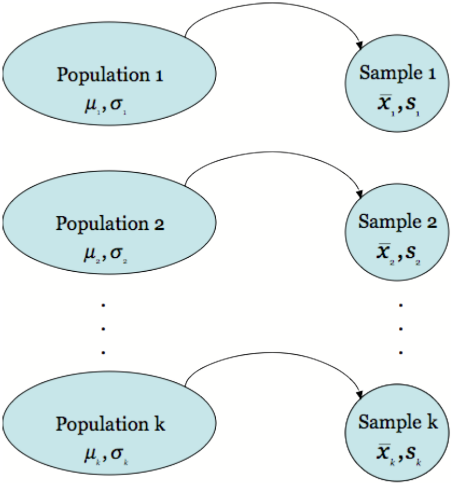
# The Basic Idea of ANOVA

The basic idea is to compare measures of variability, both between the groups and within each group, as a way to assess how different the groups really are.

*Lock5, p.540*

# Analysis of Variance Sampling Model

Draw simple random samples from $k$ independent populations to compare population means $\mu_1$, $\mu_2$, ..., and $\mu_k$:




Condition: $\sigma_1 = \sigma_2 = \cdots = \sigma_k = \sigma$

# Conditions for Applying ANOVA

1. We have $k$ independent SRSs, one from each of $k$ populations.
2. Each of the $k$ populations has a Normal distribution with an unknown mean $\mu_i$.
3. All of the populations have the same standard deviation $\sigma$, whose value is unknown.

# Case Study

## Diet Restriction and Longevity

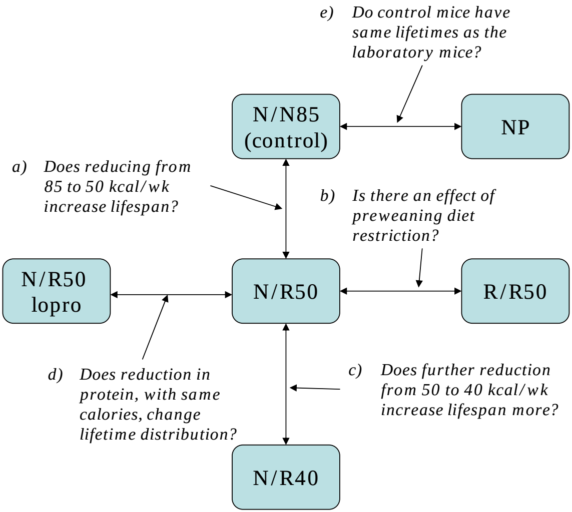
- Study of the effect of restricting caloric intake on life expectancy.
- Female mice randomly assigned to one of 6 treatment groups:
  1. NP: unlimited nonpurified standard diet for laboratory mice.
  2. N/N85 (control group): fed normally both before and after weaning. Caloric intake controlled at 85 kcal/wk after weaning.
  3. N/R50: normal diet before weaning and reduced-calorie diet of 50 kcal/wk after weaning.
  4. R/R50: reduced-calorie diet of 50 kcal/wk before and after weaning.
  5. N/R50 lopro: normal diet before weaning, restricted diet of 50 kcal/wk after weaning with dietary protein content decreased with advancing age.
  6. N/R40: normal diet before weaning, severely reduced-calorie diet of 50 kcal/wk after weaning.
- Several questions of interest to be addressed.

# Case Study

## Diet Restriction and Longevity: Study Citation

Weindruch R, Walford RL, Fligiel S, Guthrie D. The retardation of aging in mice by dietary restriction: longevity, cancer, immunity and lifetime energy intake. *J Nutr.* 1986;116(4):641-654. doi:10.1093/jn/116.4.641

# Planned Comparisons Among Groups in the Diet Restriction Study



# Diet Restriction Study

## Exploratory Plots

Use R to create plots to explore the possibility of differences among the means.

# The Big Picture of ANOVA

- Key question: Is variability across populations greater than variability within populations
  - Variability between vs within
- Hypothesis test:
  - $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ versus
  - $H_a$ : At least one $\mu_i \neq \mu_j$
- Test statistic ($F$): ratio of the variability among group (or treatment) means over the variability within samples.

$$F = \frac{\text{variation among the sample means}}{\text{variation among individuals in the same sample}}$$

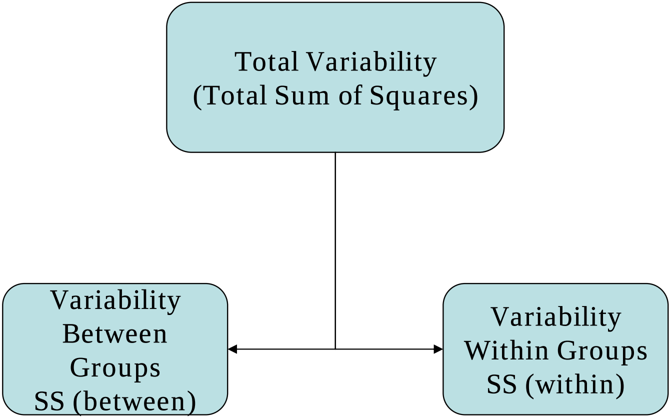
  - Large test statistic $\Rightarrow$ evidence against $H_0$
- Results are summarized in the ANOVA table.

# Notation

Key notation used in calculations for comparing variability between groups and within groups:

$$
\begin{aligned}
k &= \text{number of groups} \\
n_i &= \text{sample size for group } i \\
\bar{x}_i &= \text{sample mean for group } i \\
s_i &= \text{standard deviation for group } i \\
n &= \sum_{i=1}^{k} n_i = \text{total sample size} \\
\bar{x} &= \frac{\sum_{i=1}^{k} n_i \bar{x}_i}{n} = \text{overall mean.}
\end{aligned}
$$

# Apportioning Variability



# Variability within Groups

Measure variability within groups by the sum of squared deviations from the group mean:

$$\mathrm{SSE} = (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_k - 1)s_k^2$$

# Degrees of Freedom

- The degrees of freedom within samples is the sum of degrees of freedom for each sample.

$$\begin{aligned} \text{df(error)} &= (n_1 - 1) + (n_2 - 1) + \cdots + (n_k - 1) \\ &= n - k \end{aligned}$$

- This is also equal to the total sample size minus the number of groups.

# Mean Square within Groups (Error)

In ANOVA, the mean square measures variability as the ratio of sum of squares to degrees of freedom.

$$\begin{aligned}
\text{MSE} = \text{MS(error)} &= \frac{\text{SSE}}{\text{df(error)}} \\
&= \frac{(n_1 - 1)s_1^2 + \cdot + (n_I - 1)s_k^2}{n - k}
\end{aligned}$$

- The pooled standard deviation ($s_p$) estimates the common group-specific standard deviation:

$$s_p = \sqrt{\text{MS(error)}}$$

# Variability between Groups

Measure variability between groups by SS(between), a weighted sum of squared deviations of group means from the grand mean:

$$\text{SSG} = n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2 + \cdots + n_k(\bar{x}_k - \bar{x})^2$$

The corresponding degrees of freedom between groups is one less than the number of groups:

$$\text{df(groups)} = k - 1$$

# Mean Square between Groups

The mean square between groups, MSG or MS (treatment) is the ratio of the sum of squares over the degrees of freedom:

$$\begin{aligned} \text{MSG} &= \frac{\text{SSG}}{\text{df(groups)}} \\ &= \frac{n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2 + \cdots + n_k(\bar{x}_k - \bar{x})^2}{k-1} \end{aligned}$$

# The $F$ Statistic

- The $F$ statistic is the ratio of the mean square between to the mean square within.

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

- If the null hypothesis is true, $F$ has an $F$ distribution with $k-1$ and $n-k$ degrees of freedom
- Use the StatKey $F$ distribution web applet to find $p$-values when necessary or obtain them from R output for ANOVA using the **lm()** command we also used for regression.

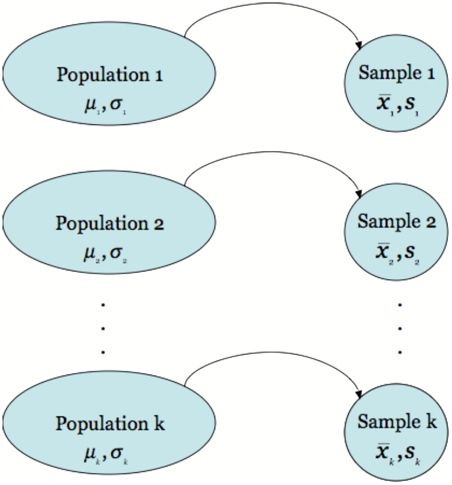
# Diet Restriction Study

See R Commands for ANOVA for R help.

# Checking Conditions

# Analysis of Variance Sampling Model

Draw samples from $k$ independent populations to compare population means $\mu_1$, $\mu_2$, . . ., and $\mu_k$:




Condition: $\sigma_1 = \sigma_2 = \cdots = \sigma_k = \sigma$

# Standard Conditions for ANOVA

- Design conditions:
  - Random samples: reasonable to consider observations a random sample from respective populations
  - Independent samples: the $k$ samples are independent of each other
- Population:
  - Normal: Population distributions are normal (not crucial if $n_i$ are large and similar)
  - Equal standard deviations:

$$\sigma_1 = \sigma_2 = \cdots = \sigma_k = \sigma$$

# Rules of Thumb for Checking ANOVA Conditions

- Inference requires random samples.
- Outliers are always problematic! Plot your data.
- ANOVA methods are robust if group samples sizes are similar and not too small.
- Ratio of largest sample SD to smallest should not be much greater than 2.
  - Biggest problem is if sample sizes are unequal and SD from a small sample is much larger than others.
- Normality is not critical if sample sizes $n_i$ are large and approximately equal.