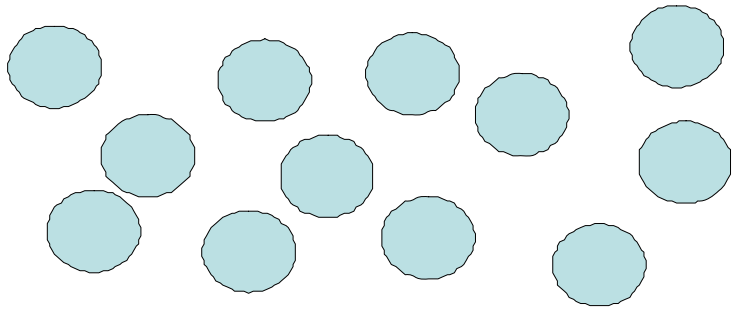
# One-way Classifications of Data and Completely Randomized Designs

A "one-way classification" of data refers to data sets that are grouped according to one criterion. It can result from designed experiments, sample surveys, or observational studies.

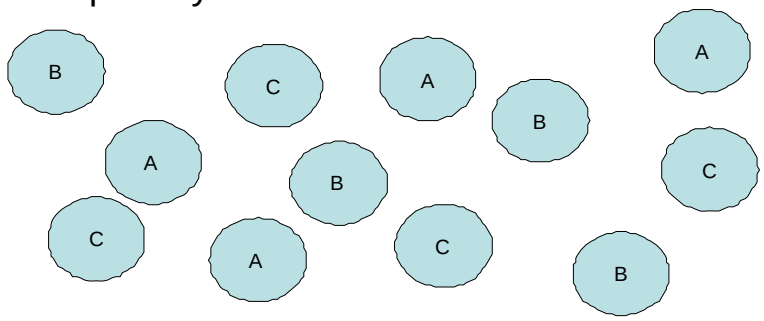
## Concepts for Completely Randomized Design

- Homogeneous EU



# Treatment Assignment in CRD

- Homogeneous EU
- Completely Randomize Treatments



**Examples of one-way classifications of data:**

Bio-availability of dietary zinc fed to sheep: Sheep were randomly assigned to diets containing supplemental zinc. Three sheep were assigned to each of eight diets. This is a completely randomized design with treatments given by the diets and sheep being the experimental units. Zinc uptake in bone was measured on each sheep.

Wear due to friction applied to samples of wood veneer material: Five brands of synthetic wood veneer material that are used for counter tops were compared for their durability. Four samples were used from each brand. The samples were subjected to a friction test in a randomly assigned order. Amounts of wear resulting from the friction test were measured on each sample. Although there are no "treatments" per se in this experiment, the brands are treated as such. The random assignment of samples to the friction test avoids systematic bias that might result from the first to the last test.

Serum zinc in dogs at the UF College of Veterinary Medicine: Blood samples were obtained from dogs that were taken to the clinic. The dogs were diagnosed according to five classifications as related to skin diseases; allergic, non-allergic, sick for non-skin disease, immune deficient, and healthy. Numbers of dogs varied in the various diagnoses. Serum zinc was measured in each of the serum samples.

Muzzle velocity of bullets: Rifle cartridges were made using three types of gunpowder. Four cartridges of each powder type were fired from a rifle in random order and muzzle velocities were measured.

Each of these examples results in data classified according to one criterion (diet, brand, diagnosis, and type, respectively). Only the first involves actual treatments, but the classification criteria are often generically referred to as "treatments."

The primary objective in each of the examples is to compare the means of the various groups. Analysis of variance provides the statistical machinery for making such comparisons.

Following are data from the synthetic veneer experiment, with summary statistics for each brand:

| Brand | ACME | AJAX | CHAMP | TUFFY | XTRA |
|---|---|---|---|---|---|
| | 2.3 | 2.2 | 2.2 | 2.4 | 2.3 |
| | 2.1 | 2.0 | 2.3 | 2.7 | 2.5 |
| | 2.4 | 1.9 | 2.4 | 2.6 | 2.3 |
| | 2.5 | 2.1 | 2.6 | 2.7 | 2.4 |
| Mean | 2.325 | 2.050 | 2.375 | 2.600 | 2.375 |
| Std. Dev. | 0.171 | 0.129 | 0.171 | 0.141 | 0.096 |
| Variance | 0.0292 | 0.0166 | 0.0292 | 0.0199 | 0.0092 |

These data constitute samples from five populations, corresponding to the brands. The means of the data from the brands differ numerically. The differences could reflect both true differences between the population means and random variation. Analysis of variance partitions variability in the data into sources of variation due to differences *Between* brands, and to differences *Within* brands. Results are usually presented in a table such as the following:

| Source of Variation | df | SS | MS | F |
|---|---|---|---|---|
| Between Brands | 4 | 0.617 | 0.154 | 7.40 |
| Within Brands | 15 | 0.315 | 0.021 | |

Details of the computations will come later. Now we present the interpretation and conclusions of the ANOVA.

The "mean square" for "Between Brands," equal to 0.154, is simply the computed variance of the brand means, multiplied by a constant. This is the summary measure of differences between the brands.

The "mean square" for "Within Brands," equal to 0.021, is the pooled variance within the brands. This is a summary measure of random variability within brands.

The ratio of mean squares, F=0.154/0.021=7.40, reflects the magnitude of variation between brands relative to random variation. This says that the measure of variation between brands is 7.40 times larger than it would be if the population means were equal. The statistical significance of an F ratio this large is p=0.0017. This is the probability of getting a value as large as 7.40 from an F distribution with 4 numerator degrees of freedom and 15 denominator degrees of freedom. Thus, we infer that there are statistically significant differences (at the p=0.0017 level) among the population means. In other words, we reject the null hypothesis that the population means are equal.

**Analysis of variance computations.** Results of an analysis of variance are usually displayed in a tabular form such as

| Source of Variation | df | SS | MS | F |
|---|---|---|---|---|
| Treatments | t - 1 | SS(T) | MS(T) | MS(T)/MS(E) |
| Error | t(n -1) | SS(E) | MS(E) | |

The term "treatment" is used generically to stand for a grouping variable, such as brands in the previous example. Likewise, "error" is a generic expression for random variation.

The primary computations are the sums or squares, SS(T) and SS(E). Let $y_{ij}$ denote an observed value, specifically, the *jth* observation in the *ith* group. Assume there are *n* observations in each group. Also, let $\bar{y}_{i.}$ denote the *mean* of the data in group *i*, and let $\bar{y}_{..}$ denote the overall mean. Likewise, let $y_{i.}$ denote the *total* for the data in group *i*, and let $y_{..}$ denote the overall total.

The SS computations can be represented in several different forms. The *definition* form illustrates the aspect of variation that is measured by each computation. First, note that the deviation of any observed value from the overall mean can be written as the sum of the deviation of the observation from its group mean plus the deviation of the group mean from the overall mean,

$$y_{ij} - \bar{y}_{..} = y_{ij} - \bar{y}_{i.} + \bar{y}_{i.} - \bar{y}_{..}$$

The sums of squares expresses a similar decomposition in terms of squared deviations.

Sums of squares, definitional form:

$$SS(T) = n\Sigma_i (\bar{y}_{i.} - \bar{y}_{..})^2$$
$$SS(E) = \Sigma_{ij} (y_{ij} - \bar{y}_{i.})^2$$
$$SS(Tot) = \Sigma_{ij} (y_{ij} - \bar{y}_{..})^2$$

The definitional form shows that SS(T) measures variation of the group means about the overall mean, SS(E) measures variation of data values about the group means, and SS(Tot) measures variation of individual data about the overall mean.

The definitional form also reveals the degrees of freedom for the sums of squares. SS(T) is a sum of squared deviations of $t$ numbers about their mean, so it has $t-1$ df. SS(E) is the sum of $t$ SS terms, each of which has $n-1$ df. Finally, SS(Tot) is the sum of squares of $nt$ observations, and has $nt-1$ degrees of freedom.

While the definitional forms illustrate the concepts of variation that are measured by the sums of squares, they are not the most convenient for computation. The sums of squares can also be expressed in terms of squared means or in terms of squares totals. Both of these expressions are more convenient for computation.

Sums of squares, computational form in terms of means:

$$SS(T) = n\Sigma_i \bar{y}_{i.}^2 - nt\bar{y}_{..}^2$$
$$SS(E) = \Sigma_{ij} y_{ij}^2 - \Sigma_i n\bar{y}_{i.}^2$$
$$SS(Tot) = \Sigma_{ij} y_{ij}^2 - nt\bar{y}_{..}^2$$

Note the each *squared mean* is *multiplied* times the number of observations in the mean.

Sums of squares, computational form in terms of totals:

$$SS(T) = \Sigma_i y_{i.}^2 / n - y_{..}^2 / nt$$
$$SS(E) = \Sigma_{ij} y_{ij}^2 - \Sigma_i y_{i.}^2 / n$$
$$SS(Tot) = \Sigma_{ij} y_{ij}^2 - y_{..}^2 / nt$$

Each *squared total* is *divided* by the number of observations in the total.

Using either of the computational forms, it is easy to see that

$$SS(Tot) = SS(T) + SS(E)$$

Also,

$$df(Tot) = df(T) + df(E)$$

The muzzle velocity example will be used to illustrate the computations. Here are the data:

| Powder | Muzzle Velocities | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| A | 27.1 | 28.1 | 27.4 | 27.7 | 28.0 | 28.1 | 27.4 | 27.1 |
| B | 28.3 | 27.9 | 28.1 | 28.3 | 27.9 | 27.6 | 28.5 | 27.9 |
| C | 28.4 | 28.9 | 28.3 | 27.9 | 28.2 | 28.9 | 28.8 | 27.7 |

The numbers of observations and means and totals are:

| Powder type | n | Totals | Means |
|---|---|---|---|
| A | 8 | 220.9 | 27.6125 |
| B | 8 | 224.5 | 28.0625 |
| C | 8 | 227.1 | 28.3875 |
| Combined | 24 | 672.5 | 28.0208 |

Sum of Squares Between Powders:

$$SS(T) = 8(27.6125 - 28.0208)^2 + 8(28.0625 - 28.0208)^2 + 8(28.3875 - 28.0208)^2$$

$$= 220.9^2/8 + 224.5^2/8 + 227.1^2/8 - 672.5^2/24$$
$$= 2.42333$$

Sum of Squares Within Powders:

$$SS(E) = (27.1 - 27.6125)^2 + \ldots + (27.1 - 27.6125)^2 + (28.3 - 28.0625)^2 + \ldots + (27.9 - 28.0625)^2 + (28.4 - 28.3875)^2 + \ldots + (27.7 - 28.3875)^2$$

$$= 27.1^2 + 28.1^2 + \ldots + 27.1^2 - 220.9^2/8 + 28.3^2 + 27.9^2 + \ldots + 27.9^2 - 224.5^2/8 + 28.4^2 + 28.9^2 + \ldots + 27.7^2 - 227.1^2/8$$

$$= 3.29625$$

$SS(Tot) = (27.1 - 28.0208)^2 + \ldots + (27.7 - 28.0208)^2$

$= 27.1^2 + 28.1^2 + 27.7^2 - 672.5^2/24$
$= 5.71958$

Analysis of Variance Table

| Source of Variation | DF | SS | MS | F | P |
|---|---|---|---|---|---|
| Between | 2 | 2.42333 | 1.21167 | 7.72 | 0.0031 |
| Within | 21 | 3.29625 | 0.15696 | | |
| Total | 23 | 5.71958 | | | |

The conclusion from this ANOVA table is the mean muzzle velocities are statistically different at the significance level p=0.0031.

**How analysis of variance works.** The principles that make ANOVA work can be understood in terms of a mathematical *model* for the data. The model is an equation that represents the data in terms of populations means and random deviations from population means.

Model for one-way classification. Let $\mu_i$ denote the population mean for the *ith* group. An observation $y_{ij}$ from this population can be expressed as the population mean pulse a random deviation from the mean,

$$y_{ij} = \mu_i + \epsilon_{ij}$$

In addition, the population mean $\mu_i$ for the *ith* group can be represented in as $\mu_i = \mu + \alpha_i$, where

$$\mu = (\Sigma_i \mu_i)/t$$

and

$$\alpha_i = \mu_i - \mu.$$

The quantity $\alpha_i$ is called the *group effect* because it represents the amount by which the population mean for group *i* differs from the average of all the population means.

Putting the parts together,

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}\ .$$

This equation represents the observation as the sum of an *overall* mean plus the *effect* of the group from which the observation was obtained, plus a *random deviation* of the observation from the group population mean.

The random deviations $\epsilon_{ij}$ are assumed to be normally distributed with mean 0 and $\sigma^2$.

*Expected mean squares* are used to motivate tests of hypothesis about population means.

The null hypothesis that the population means are equal is

$H_0$: $\mu_1 = \mu_2 = ... = \mu_t$.

It can be shown that:

MS(T) estimates $\sigma^2 + n\Sigma_i(\mu_i - \mu)^2$

and

MS(E) estimates $\sigma^2$.

The quantity $n\Sigma_i(\mu_i - \mu)^2$ would be equal to 0 if $H_0$: $\mu_1 = \mu_2 = ... = \mu_t$ is true. Thus, if $H_0$: $\mu_1 = \mu_2 = ... = \mu_t$ is true, then MS(T) would estimate $\sigma^2$. Thus, the ratio F = MS(Trt)/MS(Error) gives an indication of whether $H_0$ is true or false. If F is large, this indicates $H_0$ is false. If F is small (around 1) this indicates $H_0$ is true.

# SAS Program for Analysis of Variance of Bullets Data

```
data bullets;
  input powder $ velocity;
datalines;
A 27.1
A 28.1
A 27.4
A 27.7
A 28.0
A 28.1
A 27.4
A 27.1
B 28.3
B 27.9
B 28.1
B 28.3
B 27.9
B 27.6
B 28.5
B 27.9
C 28.4
C 28.9
C 28.3
C 27.9
C 28.2
C 28.9
C 28.8
C 27.7
;
title1 'Analysis of Variance for One-Way Classification';
proc print data=bullets;
run;
proc means data=bullets;
class powder;
run;
proc univariate data=bullets plots;
by powder;
proc glm data=bullets; class powder;
  model velocity=powder;
run;
```

Analysis of Variance for One-Way Classification

| Obs | powder | velocity |
|---|---|---|
| 1 | A | 27.1 |
| 2 | A | 28.1 |
| 3 | A | 27.4 |
| 4 | A | 27.7 |
| 5 | A | 28.0 |
| 6 | A | 28.1 |
| 7 | A | 27.4 |
| 8 | A | 27.1 |
| 9 | B | 28.3 |
| 10 | B | 27.9 |
| 11 | B | 28.1 |
| 12 | B | 28.3 |
| 13 | B | 27.9 |
| 14 | B | 27.6 |
| 15 | B | 28.5 |
| 16 | B | 27.9 |
| 17 | C | 28.4 |
| 18 | C | 28.9 |
| 19 | C | 28.3 |
| 20 | C | 27.9 |
| 21 | C | 28.2 |
| 22 | C | 28.9 |
| 23 | C | 28.8 |
| 24 | C | 27.7 |

Analysis of Variance for One-Way Classification

The MEANS Procedure

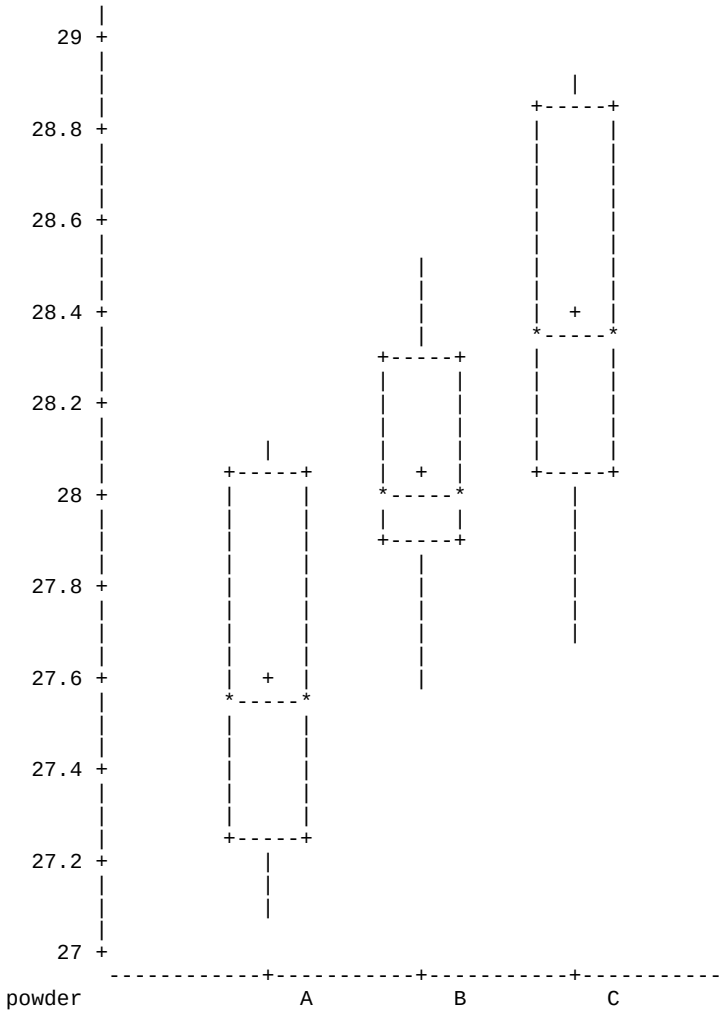
Analysis Variable : velocity

| powder | N | Mean | Std Dev | Minimum | Maximum |
|---|---|---|---|---|---|
| A | 8 | 27.6125000 | 0.4223658 | 27.1000000 | 28.1000000 |
| B | 8 | 28.0625000 | 0.2924649 | 27.6000000 | 28.5000000 |
| C | 8 | 28.3875000 | 0.4549333 | 27.7000000 | 28.9000000 |

Analysis of Variance for One-Way Classification

The UNIVARIATE Procedure
Variable: velocity

Schematic Plots



Analysis of Variance for One-Way Classification

The GLM Procedure

Class Level Information

| Class | Levels | Values |
|---|---|---|
| powder | 3 | A B C |

Number of observations 24

Analysis of Variance for One-Way Classification

The GLM Procedure

Dependent Variable: velocity

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 2 | 2.42333333 | 1.21166667 | 7.72 | 0.0031 |
| Error | 21 | 3.29625000 | 0.15696429 | | |
| Corrected Total | 23 | 5.71958333 | | | |

| R-Square | Coeff Var | Root MSE | velocity Mean |
|---|---|---|---|
| 0.423691 | 1.413902 | 0.396187 | 28.02083 |

| Source | DF | Type I SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| powder | 2 | 2.42333333 | 1.21166667 | 7.72 | 0.0031 |

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| powder | 2 | 2.42333333 | 1.21166667 | 7.72 | 0.0031 |