How to randomize an experiment in R. Main point: Use the computer to randomize units to treatments.

Illustration of R function `sample`: uses a random number generator

```
committee <- c("Al", "Bob", "Cathy", "Don", "Ed", "Gail")
set.seed(1)
subcommittee <- sample(x=committee, size=3, replace=F)
subcommittee
[1] "Bob"   "Gail"  "Cathy"
```

Set the starting point for pseudo r.n.g. to be a seed of 1

From the committee, select 3 people at random and w/o replacement.

```
individuals <- 1:20
set.seed(1); sample(x=individuals, size=10, replace=F)
[1]  6  8 11 16  4 14 15  9 19  1
```

This randomly splits the individuals into two groups, each of size 10.

How to randomize an experiment in R (Completely randomized design)

Example Obtain a randomization of units to treatments for a CRD with four treatments and two replicates. Use R function `sample`

Solution:

Label the eight units: $1, 2, 3, 4, 5, 6, 7, 8$

List the treatments: $T1, T1, T2, T2, T3, T3, T4, T4$

Get a random permutation of the units, and assign this permutation of the units to the treatments as listed above:

```
> set.seed(12)
> sample(1:8,replace=FALSE)
[1] 1 6 7 2 8 4 3 5
```

Randomization scheme:

Table. Randomization scheme

Treatment	Unit ids
1	1, 6
2	2, 7
3	4, 8
4	3, 5

One-way Analysis of Variance (ANOVA)

The basic setup is simple: We have several (≥ 3) groups. This is like the two-sample problem, but with at least one additional group. The groups are *independent* (separate and unrelated or unpaired).

A grouping factor can be *experimental*, perhaps a control group and two or more treatment groups, e.g. two different doses of a medicine, or two different medicines.

Or, a grouping factor can be *observational*, not assigned by the experimenter but a feature of the data.

Example 1. Dose of medicine What is the best dosage level of a particular medicine? In order to compare effectiveness of three dosage levels, first recruit 30 patients with the medical problem to participate in a pilot study. Assign each patient at random to one of the three drug dosage levels, such that exactly ten patients receive each dosage level. One of the groups is a placebo group (dosage is zero, and a fake medicine is given). This is a *balanced* completely randomized one-way experimental design.

Example 2. Absorptive Properties of Paper Towels In an experiment to compare paper towel brands, five sheets of paper were selected from each of Bounty, Scott, Viva, and Generic. Twenty 6-ounce beakers of water were prepared, and the twenty paper towel sheets were randomly assigned to the beakers. This is a balanced CR one-way design with one factor at four levels.

In both of the examples, we commonly view the predictor variable or factor as a single, qualitative predictor variable. [In ANOVA, we will allow a different mean parameter for each level of the factor.]

The resulting model can be viewed as a ~~regression~~ general linear model with several dummy variable predictors.

For r levels of the factor, we need $r - 1$ dummy variables. Spell this out:

Example. Find best dosage of blood pressure medicine Factor is dose, levels Placebo, Low, High (corresponding to quantitative doses of 0, 20, and 40 mg, say)

Define dummy variables as follows:
Let $X_1 = 1$ for Placebo, 0 otherwise
Let $X_2 = 1$ for Low, 0 otherwise

$$EY = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

Group	Mean (GLM)	ANOVA parameter
Placebo	$\beta_0 + \beta_1$	$\mu_{Placebo}$
Low	$\beta_0 + \beta_2$	μ_{Low}
High	β_0	μ_{High}

Choice between ANOVA model and regression model

use dose as a single quantitative predictor

Example 1. Effectiveness of Medicine Dose The goal is to compare effectiveness of various doses of a certain medicine in reducing blood pressure.

- Experimental units: 30 subjects
- Response variable: Y is change in blood pressure (pre - post)
- One factor, dose of medicine. Let's say there are ~~six~~ 5 levels. (We need multiple doses for the purposes of this discussion.)
- Design: Randomly assign the 30 subjects to the ~~six~~ 5 groups. This is a CRD, with one factor at ~~six~~ 5 levels.
- Analysis: You can use one-way analysis of variance. OR: use regression analysis. (Regression - Simple - $EY = \beta_0 + \beta_1 X$, dose)

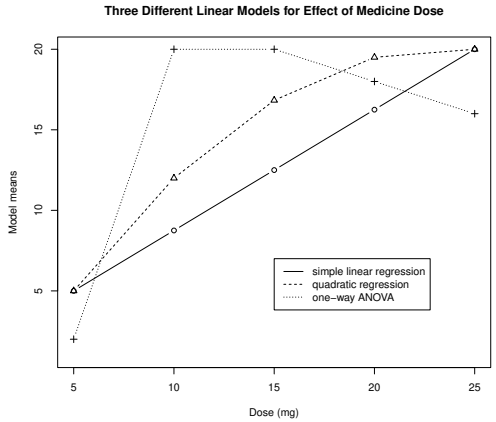
Because the predictor variable X is quantitative, there are two possible approaches to the analysis.

The reason this is important to think about is that for dose to be treated as numerical, you'd need a regression model to be appropriate.

Ex. 1 Effectiveness of Medicine Dose, con.

Compare the following models for this situation: simple linear regression, quadratic regression, and one-way ANOVA:

$EY = \beta_0 + \beta_1 X + \beta_2 X^2$



Model 1, simple linear regression: $EY_i = \beta_0 + \beta_1 x_i$ Key assumption: The slope is constant over the whole range of doses. This might not hold.
Model 2, quadratic regression: $EY_i = \beta_0 + \beta_1 x_i + \beta_{11} x_i^2$ (we would center the x's first to do quadratic regression right) This model allows the means to increase less as doses increase and is sometimes a realistic model.
Model 3, one-way ANOVA: This is most general model of the three: allows arbitrarily different means for each dose. This figure has only one out of a huge number of possible patterns of the means.

Example for Illustration of Statistical Analysis

Example. Kenton Food Company (Section 16.4, p. 691) The company wants to compare four different package designs for a new breakfast cereal in terms of sales.

- Experimental units: 20 stores [2]
- Data: A fire occurred in one store, which had to be dropped from the study. Response variable Y was sales, in number of cases.
- Design: Completely randomized design (CRD) with package design as the single factor, with four levels.
- Further points: Shelf space, level of advertising, and other factors which might affect sales were kept constant among the stores in the study.

Key point about the model The model says there could be arbitrarily different means μ_A, μ_B, μ_C, μ_D in the four groups.

[2] comparable in sales volume and location

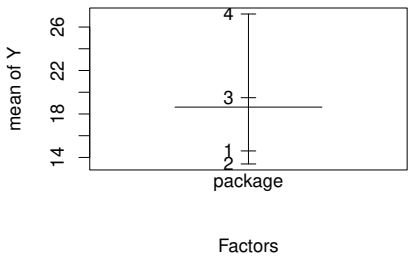
The Kenton Food Company wished to test four different package designs for a new breakfast cereal. Twenty stores, with approximately equal sales volumes, were selected as the experimental units. Each store was randomly assigned one of the package designs, with each package design assigned to five stores. A fire occurred in one store during the study period, so this store had to be dropped from the study. Hence, one of the designs was tested in only four stores. The stores were chosen to be comparable in location and sales volume. Other relevant conditions that could affect sales, such as price, amount and location of shelf space, and special promotional efforts, were kept the same for all of the stores in the experiment. Sales, in number of cases, were observed for the study period, and in the file `CH16TA01.txt`. This study is a completely randomized design with package design as the single, four-level factor.

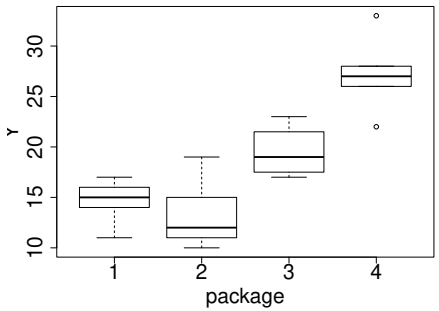
Example. Kenton Food Company. Recall: Four package designs for a breakfast cereal are being tested in five stores each (except for one package design which was tested in four stores).

Data:

	Store (j)				
Package Design	1	2	3	4	5
i	Y_{i1}	Y_{i2}	Y_{i3}	Y_{i4}	Y_{i5}
1	11	17	16	14	15
2	12	10	15	19	11
3	23	20	18	17	
4	27	33	22	26	28

Plot the data:





ALWAYS PLOT YOUR DATA

Here we have a line graph of factor level means, and comparative boxplots for the four treatment groups.
Both graphs are useful; they complement each other.

Line graph of factor level means invites the viewer to focus on the key questions—do the means differ, and if so, which means differ and how much?

The boxplots introduce variability within the groups into view, and thus show us the question we need to answer with statistical inference: Is the observed difference real, or is it just due to chance?

Reminder: A boxplot is a graphical display of the five-number summary: min, Q1, Q2, Q3, max. The "slick" version of the boxplot includes a method to detect outliers.

Consider the boxplot for Package 1. Since there are five observations, what is being plotted, exactly?

Word to the wise: For small sample size, boxplots are still useful, but don't overinterpret the boxes (the middle fifty percent of the data is only a few observations).

There are two main parts in the analysis:

1. An overall test to see if there is statistical evidence that there exist *any* differences.
2. A more detailed *follow-up* analysis to decide which of the populations differ, and to estimate how large the differences are.

Let $r =$ number of levels of the explanatory variable (number of treatment groups for example). Let $n_i =$ number of cases (experimental units) in i^{th} group, and let $n_T = \Sigma_{i=1}^{r} n_i$ be the total number of observations.

Let the population mean parameters be μ_i, $i = 1, \ldots, r$.

Hypotheses to be tested:

H_0: $\mu_1 = \mu_2 = \cdots = \mu_r$ vs.

H_a: at least two of the means are not equal.

Assumptions:

- Each of the r population distributions is normal.
- The r standard deviations are all equal.
- All n_T observations are taken independently.

Let $\mu_i =$ the mean sales volume (number of cases) that would be seen in the whole population of stores, similar to those in the study, if Package Design i was used.

Independence assumption really has two parts. We assume the r groups are independent, and that observations within each group are independent.

Alternative hypothesis is the opposite of the null hypothesis.
H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$
H_a: Not H_0
Better: H_a: For at least one pair of means μ_i, μ_j, $\mu_i \neq \mu_j$

Common error: