Comparative Experiments

E.g. Tension bond strength of mortar (kgf/cm$^2$)

Measurements of strength of 10 samples of a modified mortar formulation, and 10 samples of the unmodified formulation:

- Strengths for both formulations are broadly similar;
- On average, modified is slightly weaker;
- Is the difference real?
The data (cement.txt):

<table>
<thead>
<tr>
<th>j</th>
<th>Modified</th>
<th>Unmodified</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>16.85</td>
<td>16.62</td>
</tr>
<tr>
<td>2</td>
<td>16.40</td>
<td>16.75</td>
</tr>
<tr>
<td>3</td>
<td>17.21</td>
<td>17.37</td>
</tr>
<tr>
<td>4</td>
<td>16.35</td>
<td>17.12</td>
</tr>
<tr>
<td>5</td>
<td>16.52</td>
<td>16.98</td>
</tr>
<tr>
<td>6</td>
<td>17.04</td>
<td>16.87</td>
</tr>
<tr>
<td>7</td>
<td>16.96</td>
<td>17.34</td>
</tr>
<tr>
<td>8</td>
<td>17.15</td>
<td>17.02</td>
</tr>
<tr>
<td>9</td>
<td>16.59</td>
<td>17.08</td>
</tr>
<tr>
<td>10</td>
<td>16.57</td>
<td>17.27</td>
</tr>
</tbody>
</table>
An R session

```r
> cement <- read.table("data/cement.txt", header = TRUE)
> print(cement)

j Modified Unmodified
 1   1  16.85   16.62
 2   2  16.40   16.75
 3   3  17.21   17.37
 4   4  16.35   17.12
 5   5  16.52   16.98
 6   6  17.04   16.87
 7   7  16.96   17.34
 8   8  17.15   17.02
 9   9  16.59   17.08
10 10  16.57   17.27
```
> print(summary(cement))
  Modified   Unmodified
Min.   :16.35  Min.   :16.62
1st Qu.:16.53  1st Qu.:16.90
Median :16.72  Median :17.05
Mean   :16.76  Mean   :17.04
3rd Qu.:17.02  3rd Qu.:17.23
Max.   :17.21  Max.   :17.37

> boxplot(cement)
Comparison box plots:
We may need to convert data in this format to one where the measurements are all in one column, called a "long" versus "wide" format.

The R function `reshape` will do the conversion:

```r
cementLong <- reshape(cement, varying = 2:3, idvar = "Obs",
  v.names = "Strength", direction = "long",
  timevar = "Formulation",
  times = names(cement)[2:3])
```

The `boxplot` function also works with data in this format; we specify the plots using a `formula`, which specifies the response variable, and the factor that influences it:

```r
boxplot(Strength ~ Formulation, data = cementLong)
```
A SAS **program** and **output**:

```sas
options linesize = 80;
ods html file = 'cement.html';

data cement;
  infile 'data/cement.txt' firstobs = 2;
  input j mod unmod;

proc means data = cement mean stddev min p25 p50 p75 max;
  var mod unmod;
```
/* make a (long) dataset with a response and a factor */
data mod;
  set cement;
  form = 'mod';
  strength = mod;

data unmod;
  set cement;
  form = 'unmod';
  strength = unmod;

data byform;
  set mod unmod;

proc boxplot data = byform;
  plot strength * form;
run;
Review of Statistical Concepts

Each measurement is the observed value of a *random variable*.

Different measurements are *independent*.

Measurements in the two samples come from possibly different *populations*;

in other words, the random variables have possibly different *distributions*. 
The simplest distribution for continuous measurements is the *normal* distribution; with mean \( \mu \) and standard deviation \( \sigma \),

\[
f(y) = \frac{1}{\sqrt{2\pi \sigma^2}} e^{-\frac{(y - \mu)^2}{2\sigma^2}}.
\]
One reason that the normal distribution is often a good approximation is the *Central Limit Theorem*:

- roughly, a random variable that is the sum of many small independent contributions is approximately normally distributed.
Sampling Distributions

If $Y_1, Y_2, \ldots, Y_n$ are a random sample from the normal distribution $N(\mu, \sigma^2)$, and

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

is the sample mean and

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

is the sample variance, then:
the sampling distribution of $\bar{Y}$ is $N(\mu, \sigma^2 / n)$, or equivalently

$$\frac{\bar{Y} - \mu}{\sigma / \sqrt{n}} \sim N(0, 1);$$

the distribution of $S^2$ is

$$\frac{(n - 1)S^2}{\sigma^2} \sim \chi^2_{n-1},$$

the $\chi^2$ distribution with $n - 1$ degrees of freedom;

the ratio

$$\frac{\bar{Y} - \mu}{S / \sqrt{n}} \sim t_{n-1},$$

Student’s $t$-distribution with $n - 1$ degrees of freedom.
We use the first and third of these to make confidence intervals for $\mu$:

- if $\sigma$ is known, use the first;
- if $\sigma$ is unknown, use the third.

We use the second to find a confidence interval for $\sigma$. 
Statistical Inference

A model for the mortar strength data:

\[ y_{i,j} = \mu_i + \epsilon_{i,j}, \quad i = 1, 2, \quad j = 1, 2, \ldots, n_i, \]

where \( \epsilon_{i,j} \sim N(0, \sigma_i^2) \).

The statistical hypotheses:

Null hypothesis \( H_0 : \mu_1 = \mu_2 \)
Alternate hypothesis \( H_1 : \mu_1 \neq \mu_2 \).
How to Decide

Intuitively, we’ll reject $H_0$ if $\bar{y}_1$ and $\bar{y}_2$ are very different.

We need a test statistic:

$$t_0 = \frac{\bar{y}_1 - \bar{y}_2}{\text{estimated standard error}(\bar{y}_1 - \bar{y}_2)}.$$
$t_0$ measures the difference in means, relative to the estimated standard error of that difference: assuming $\sigma_1 = \sigma_2 = \sigma$,

\[
\text{standard error}(\bar{y}_1 - \bar{y}_2) = \sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}};
\]

we estimate $\sigma^2$ by the pooled variance

\[
S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}.
\]

So

\[
\frac{\bar{y}_1 - \bar{y}_2}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}.
\]
We find $t_0 = -2.187$.

If $H_0$ were true, $t_0$ would be $t$-distributed with $n_1 + n_2 - 2 = 18$ degrees of freedom, and from tables,

$$P(|t| > 2.101) = 0.05.$$ 

So, if $H_0$ were true, we would be unlikely to get $|t_0| > 2.101$ ($P < 0.05$).

So we reject $H_0$; the data suggest that the two formulations really do have different strengths.
In R, still assuming equal variances:

```r
> t.test(Strength ~ Formulation, cementLong, var.equal = TRUE)

Two Sample t-test

data:  Strength by Formulation
t = -2.1869, df = 18, p-value = 0.0422
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.54507339  -0.01092661
sample estimates:  
  mean in group Modified mean in group Unmodified
              16.764              17.042
```
Using the two-column version of the data:

```r
> t.test(cement$Modified, cement$Unmodified, var.equal = TRUE)

Two Sample t-test

data:  cement$Modified and cement$Unmodified
t = -2.1869, df = 18, p-value = 0.0422
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  -0.54507339 -0.01092661
sample estimates:
mean of x  mean of y
  16.764    17.042
```