

Comparative Experiments

E.g. Tension bond strength of mortar:

- measurements of strength of 10 samples of a modified mortar formulation, and 10 samples of the unmodified formulation;
- broadly similar;
- on average, modified slightly weaker;
- is the difference real?

The data (cement.txt):

j	Modified	Unmodified
1	16.85	16.62
2	16.40	16.75
3	17.21	17.37
4	16.35	17.12
5	16.52	16.98
6	17.04	16.87
7	16.96	17.34
8	17.15	17.02
9	16.59	17.08
10	16.57	17.27

An R session

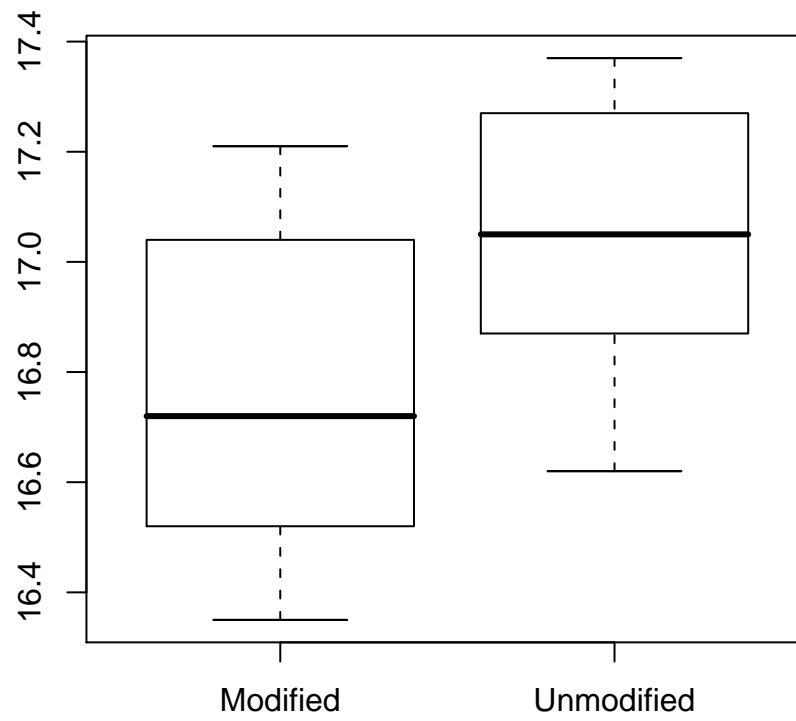
```
> cement = read.table("cement.txt", header = TRUE);  
> cement = cement[ , -1]; # drop the first column  
> print(cement);
```

	Modified	Unmodified
1	16.85	16.62
2	16.40	16.75
3	17.21	17.37
4	16.35	17.12
5	16.52	16.98
6	17.04	16.87
7	16.96	17.34
8	17.15	17.02
9	16.59	17.08
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:



A SAS program and output:

```
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 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;
```

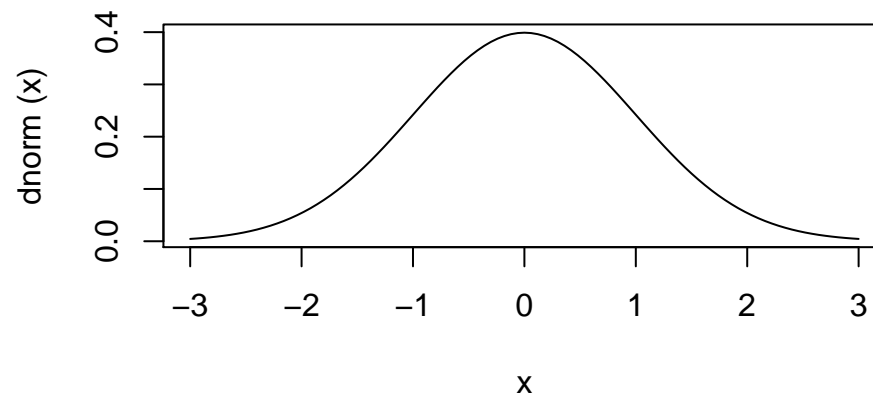
Review of Statistical Concepts

- 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 μ and standard deviation σ ,

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}}e^{-(1/2)[(y-\mu)/\sigma]^2}.$$

Standard normal: mu = 0, sigma = 1



- 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, \dots, 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

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

the χ^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 μ :
 - if σ is known, use the first;
 - if σ is unknown, use the third.
- We use the second to find a confidence interval for σ .

Statistical Inference

- A model:

$$y_{i,j} = \mu_i + \epsilon_{i,j}, \quad i = 1, 2, \quad j = 1, 2, \dots, n_i,$$

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 σ^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

$$t_0 = \frac{\bar{y}_1 - \bar{y}_2}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}.$$

- 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.