Stat 515: Introduction to Statistics

Chapter 10

Designed Studies: Vocabulary

- The **response variable** is the variable of interest to be measured in an experiment
  - Essentially this is what we would like to model or predict
- **Factors** are the variable whose effect on the response is of interest
  - Essentially this is what we would like to use to model or predict the **response**

Designed Studies: Vocabulary

- Note: Both the response(s) and factor(s) are either qualitative, quantitative discrete or quantitative continuous just as we've learned about variables earlier in the semester.

Designed Studies: Vocabulary

- **Factor Levels** are the values of the factor
  - We most often refer to the categories that make up a qualitative variable as ‘levels’
- **Treatments** of an experiment are the factor level combinations utilized

Designed Studies: Vocabulary

- An **experimental unit** is the object on which the response and factors are observed
  - These are often, but not always people

Designed Studies: Vocabulary

- Recall the difference between designed experiments and observational studies
- An **observational study** measures the response variable without attempting to influence the value of either the response or explanatory variables.
- A **designed study** occurs when a researcher assigns the individuals or subjects into groups and intentionally affects their explanatory variables (think treatments)

Example

- C. Myrray Parkes headed a study of 4,486 men of 55 years of age and older who had their wives die in 1957. For up to nine years, these widowers were tracked and 213 died during the first six months – that’s about 5%..
- This experiment is a **observational study**. C Myrray Parkes didn’t murder 4,486 women in 1957 just to do this study.

Example

- Herbet Benson, MD headed a study in 2005 to see if intercessory prayer influenced recovery from bypass surgery. There were three groups in the study: 1. Those being prayed for that didn’t know 2. Those being prayed for that did know 3. Those not being prayed for
- This is a **designed study** because the researchers assigned different patients to different groups; they controlled who was prayed for and who wasn’t instead of just observing and asking the families whether or not they had friends and families praying for the patient.

Designed Studies: Vocabulary

- Note: both of these studies were planned but we only poke and prod the experimental units in the second study where the researched assigned different “treatments” to each patient.

Designed Studies: Vocabulary

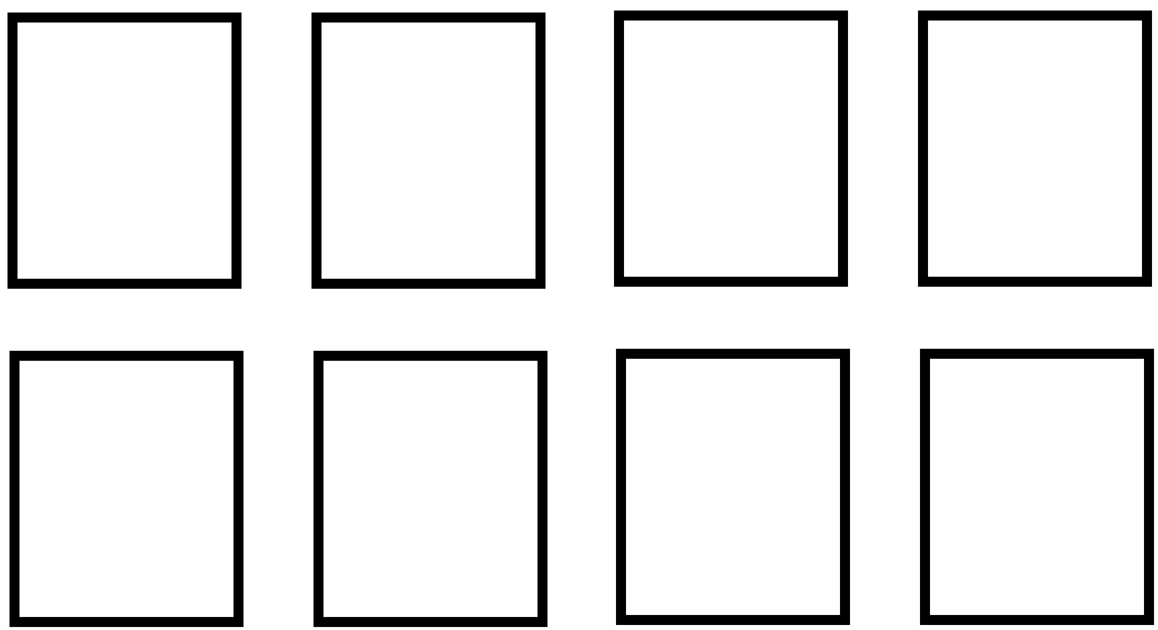
- A **completely randomized designed** is a design where the treatments are randomly assigned to the experimental units
- Note: Why we need a random sample is obvious by now from the need in Ch 7-9 to have our inference be applicable to the rest of the population – why the treatments need to be random is less obvious

Designed Studies: Vocabulary

- We call an experiment **balanced** if we assign the same number of **experimental units** to each **treatment**

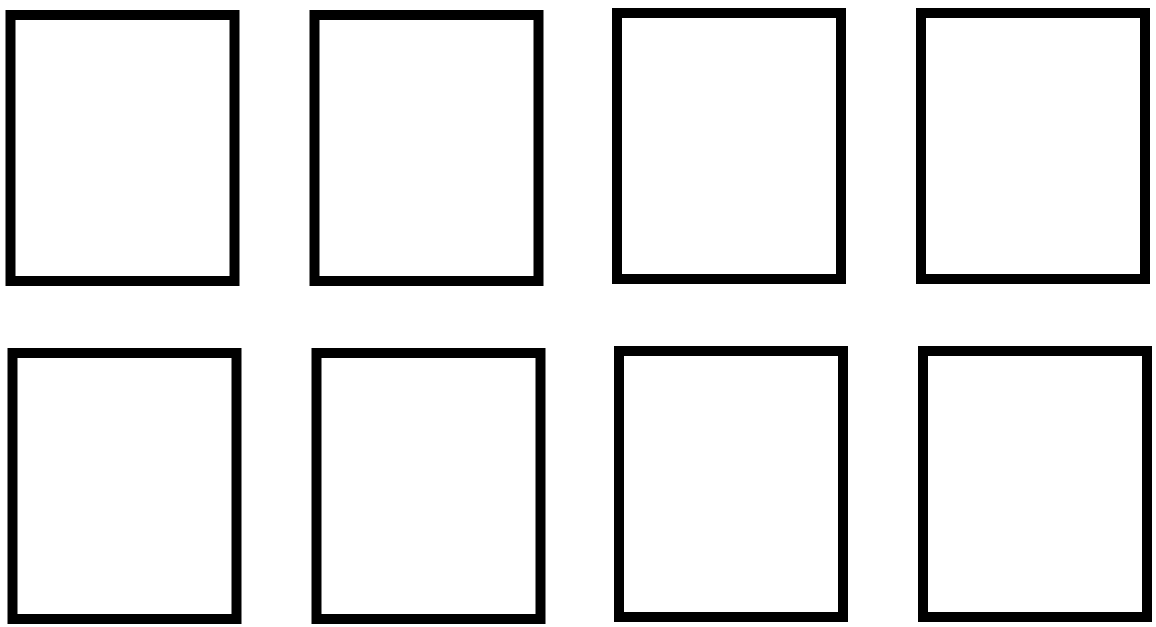
Example

- Consider a field of plots where we plant corn



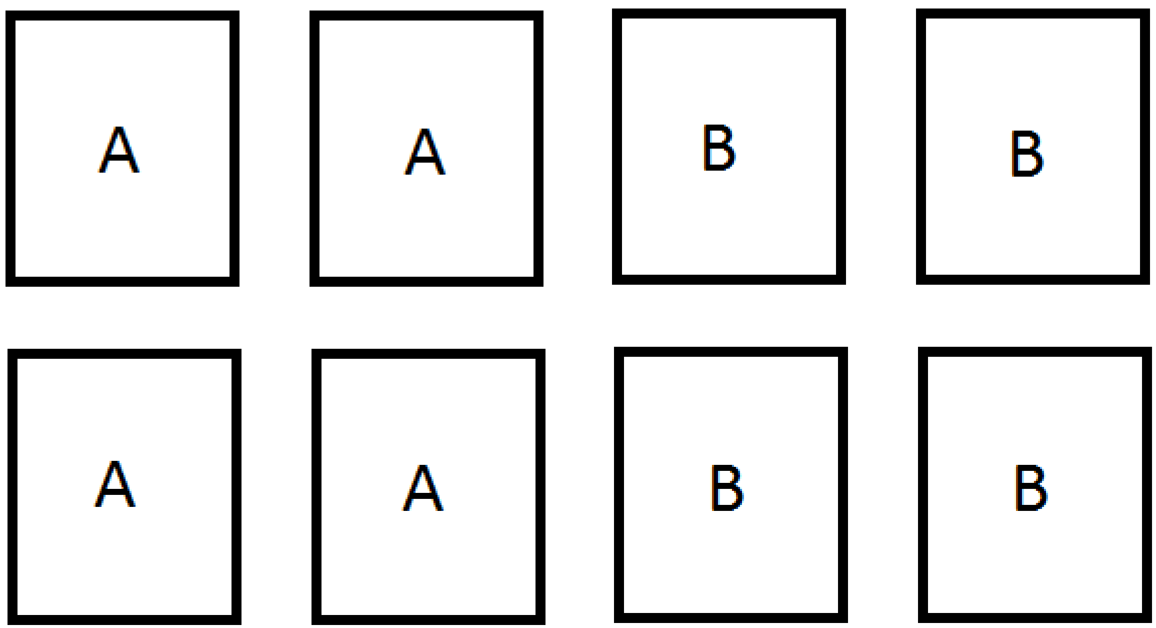
Example

- Here, we consider two different fertilizers and we want to know which works the best



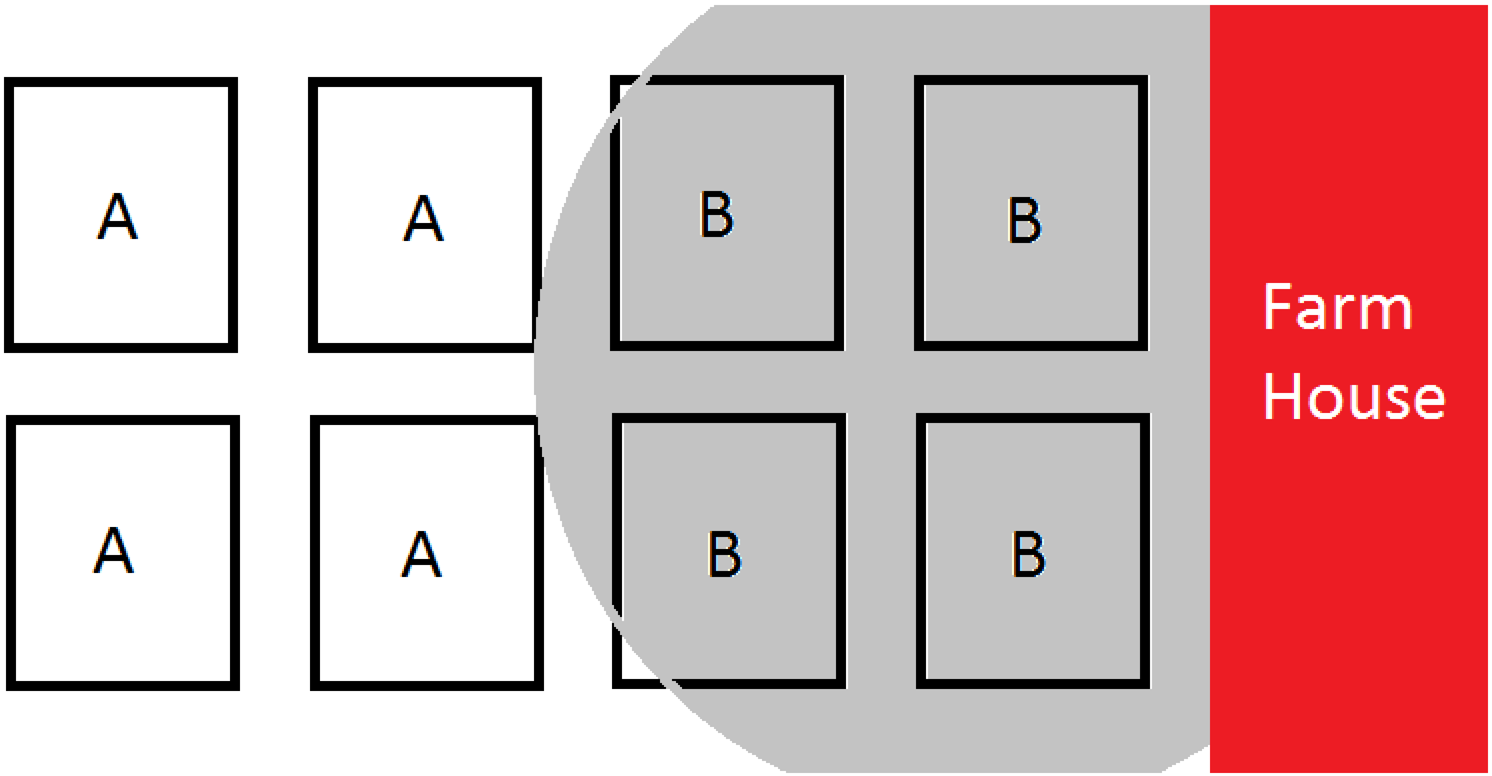
Example

- Consider this very 'un-random' assortment of fertilizers A & B, what's a possible drawback?



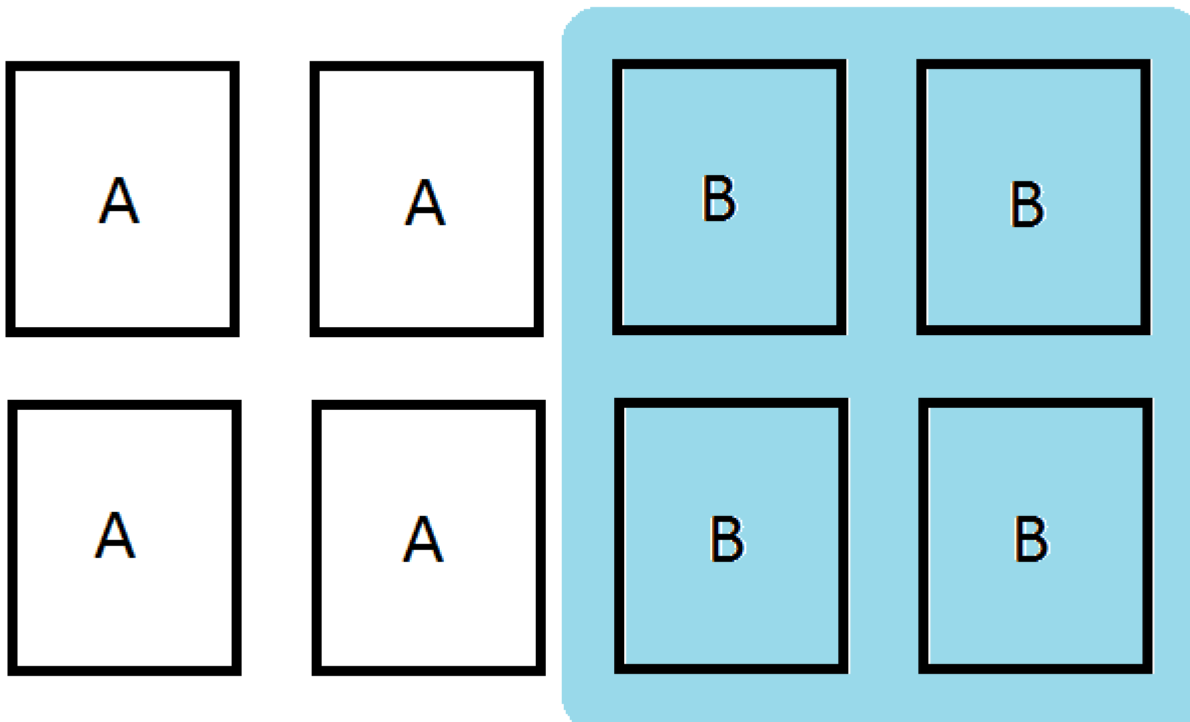
Example

- Perhaps, the farm house puts shade on the four plots to the right. In this case B would grow less, perhaps even if B is superior to A.



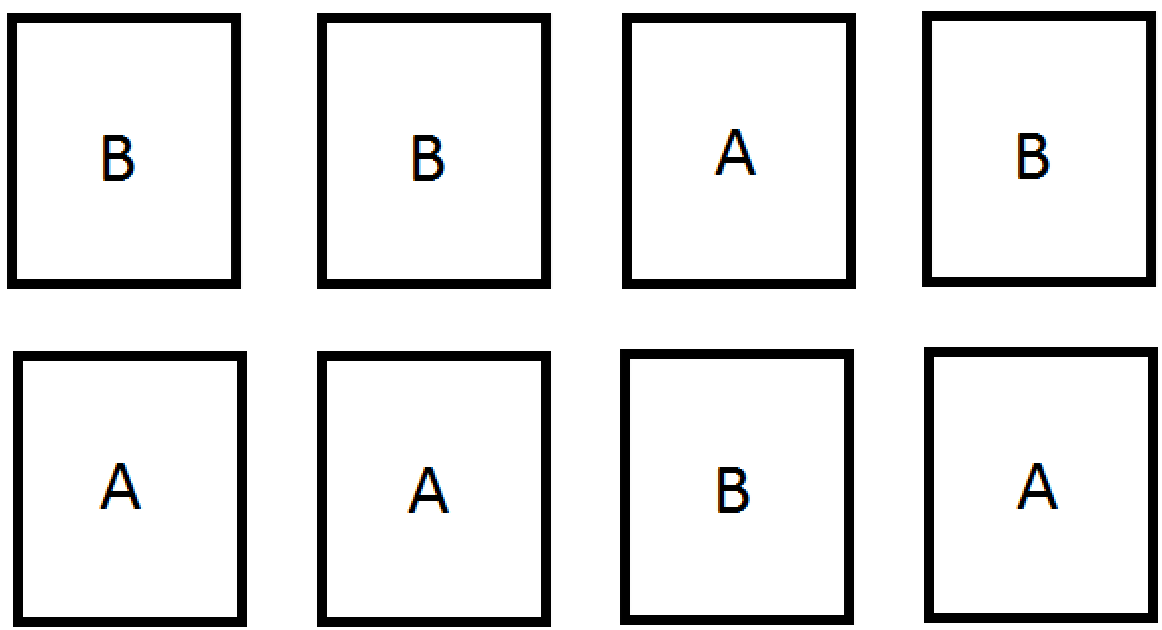
Example

- Consider a field with poor drainage suppose that the plots to the right puddled with water after rain. In this case B would grow less, perhaps even if B is superior to A



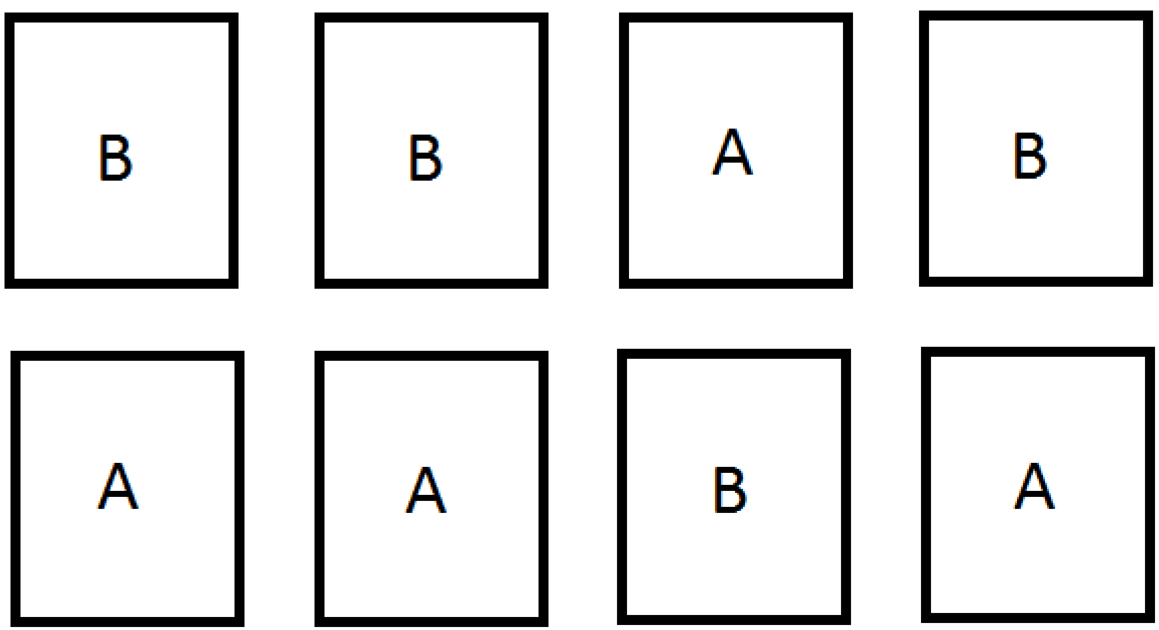
Example

- In either case, randomization would have helped us. Something like the randomization below would put A and B in both situations.



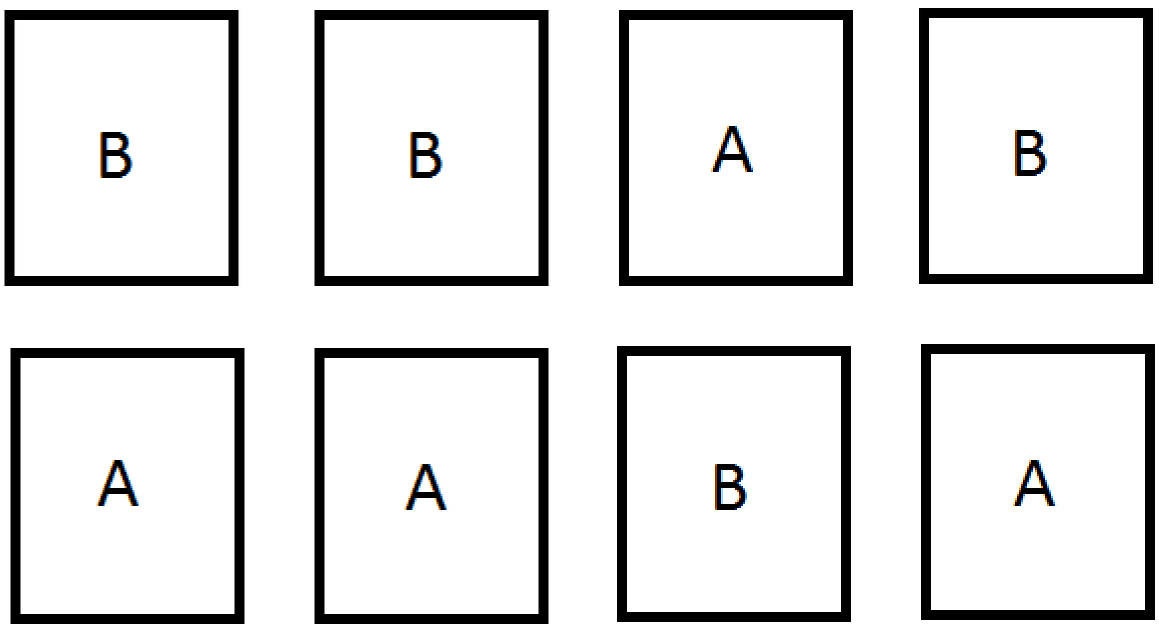
Example

- In either case, randomization would have helped us. Something like the randomization below would put A and B in both situations.



Example

- In either case, randomization would have helped us. Something like the randomization below would put A and B in both situations.



Single Factor Design

- In a **Single Factor Design** we are only considering one factor in the modeling/prediction of the response variable

- We're interested in testing

H_0: $\mu_1 = \mu_2 = \cdots = \mu_k$: The means are equal

H_A: At least one of the μ_i is different

Single Factor Design: Values of Interest

- **Sum of Squares for Treatments (SST):** the variation between the treatment means

$$SST = \sum n_i(\overline{x_i} - \bar{x})^2$$

$$n_i = \text{sample size for treatment } i$$
$$\overline{x_i} = \text{sample mean for treatment i}$$
$$\bar{x} = \text{sample mean overall}$$

Single Factor Design: Values of Interest

- **Sum of Squares for Error (SSE):** the variation within the treatments

$$SSE = \sum\left(x_{1j} - \overline{x_1}\right)^2 + \sum\left(x_{2j} - \overline{x_2}\right)^2 + \cdots + \sum\left(x_{kj} - \overline{x_k}\right)^2$$

$$x_{\mathrm{ij}} = observations\ j\ from\ treatment\ 1$$

$$\overline{x_i} = \text{sample mean from treatment i}$$

$$\bar{x} = \text{sample mean overall}$$

Single Factor Design: Values of Interest

- **Sum of Squares for Error (SSE):** the variation within the treatments

Recalling s^2:

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

$$n_i = \text{sample size for treatment } i$$
$$\mathrm{s_i^2} = \text{sample variance for treatment i}$$

Single Factor Design: Values of Interest

- **Mean Square for Treatments (MST):** measures the variation among the treatment

$$MST = \frac{SST}{k-1}$$

$$SST = sum\ of\ squares\ for\ treatments$$

$$k = \text{the number of treatments}$$

Single Factor Design: Values of Interest

- **Mean Square for Error (MSE):** measures the variation among the treatment

$$MSE = \frac{SSE}{n-k}$$

$$SS\mathrm{E} = sum\ of\ squares\ for\ Error$$

$$\mathrm{n} = \text{the number of experimental units}$$

$$k = \text{the number of treatments}$$

Single Factor Design: Values of Interest

- **F Statistic(F):** Completes the test that we're interested in

$$F = \frac{MST}{MSE}$$

$$SS\mathrm{E} = sum\ of\ squares\ for\ Error$$

$$\mathrm{n} = \text{the number of experimental units}$$

$$k = \text{the number of treatments}$$

Single Factor Design: Values of Interest

H_0: $\mu_1 = \mu_2 = \cdots = \mu_k$: The means are equal

H_A: At least one of the μ_i is different

Assumptions: 1) Samples are randomly selected

2) The k distributions are normal

3) $\sigma_1 = \sigma_2 = \cdots = \sigma_k$

Test Statistic: $F = \frac{MST}{MSE}$

Reject when: $F > F_{1-\alpha} = qf(1-\alpha, k-1, n-k)$

Example

- Suppose we have the following results for corn yield (in bushels) per plot of land with either fertilizer A or fertilizer B

B 102	B 99	A 102	B 106
A 101	A 95	B 105	A 99

Example

- $\overline{x_A} = 99.25$
- $\overline{x_B} = 103$
- $s_A^2 = 9.58333$
- $s_B^2 = 10$

Example

- In this case we're only considering **one factor**: fertilizer which has **two levels** A and B.

- We're interested in testing

H_0: $\mu_A = \mu_B$: The fertilizer work equally

H_A: $\mu_A \neq \mu_B$: The fertilizer don't work equally

Example

$$SST = \sum n_i(\overline{x_i} - \bar{x})^2$$

$$= 4(99.25 - 101.125)^2 + 4(103 - 101.125)^2$$

$$= 28.125$$

Single Factor Design: Values of Interest

$$
\begin{aligned}
SSE &= (n_A - 1)s_A^2 + (n_B - 1)s_B^2 \\
&= (4 - 1)9.58333 + (4 - 1)10 \\
&= 58.7499
\end{aligned}
$$

Single Factor Design: Values of Interest

$$MST = \frac{SST}{k-1} = \frac{28.125}{2-1} = 28.125$$

Single Factor Design: Values of Interest

$$MSE = \frac{58.7499}{8-2} = 9.79165$$

Single Factor Design: Values of Interest

$$F = \frac{MST}{MSE} = \frac{28.125}{9.79165} = 2.872345$$

Single Factor Design: Values of Interest

Reject when: $F > F_{1-\alpha} = qf(1-\alpha, k-1, n-k)$

F=2.872345

$F_{1-\alpha}$ = qf(.95,2-1,8-2)=5.987378

2.872345<5.987378 so we reject the null hypothesis in favor of the alternative.

Note: our pvalue agrees with this decision

In R

```
plotYield<-c(102,101,95,99,102,99,106,105)
fert<-c('A','A','A','A','B','B','B','B')
farming<-lm(plotYield~fert)
summary(farming)
anova(farming)
```

- This gives us the F statistic and the associated pvalue, along with other measurements

In R

summary(farming)

```
Call:
lm(formula = plotYield ~ fert)

Residuals:
   Min     1Q Median     3Q    Max 
-4.250 -1.750  0.750  2.188  3.000 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   99.250      1.565  63.435 1.03e-09 ***
fertB          3.750      2.213   1.695    0.141    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 3.129 on 6 degrees of freedom
Multiple R-squared:  0.3237,	Adjusted R-squared:  0.211 
F-statistic: 2.872 on 1 and 6 DF,  p-value: 0.141
```

Mean

Difference for B

t test for significance

F statistic

In R

anova(farming)

```
Analysis of Variance Table

Response: plotYield
          Df Sum Sq Mean Sq F value Pr(>F)
fert       1 28.125 28.1250  2.8723  0.141
Residuals  6 58.750  9.7917
```

SST

SSE

MST

MSE

F statistic

p-value