Design of Experiments:

Let us consider the following problem:

We are interested in comparing if there is a difference in the different machine learning algorithms

- Decision Trees (DT)
- Neural Networks (NN)
- Bayesian Learners (BL)

on *mapping sensor event sequences to activity labels*

to determine if there is a difference between them in terms of mean **accuracy.**

How would you design this study?

First is it important to design this study?

Let us say

- I pick 3 students from among 1st year Grad students and assign them the task of using DT algorithms.
- I pick 3 students from among 2nd year Grad students and assign them the task using of NN algorithms.
- I pick 3 students from among 3rd year Grad students and assign them the task of BL algorithms.

IS THIS A GOOD EXPERIMENT?

HOW would you like to design this differently?

Think of what you want to guard against when you design an experiment?

A major player we want to guard against is BIAS.

What is bias?

What kind of bias would you expect in the study we mentioned?

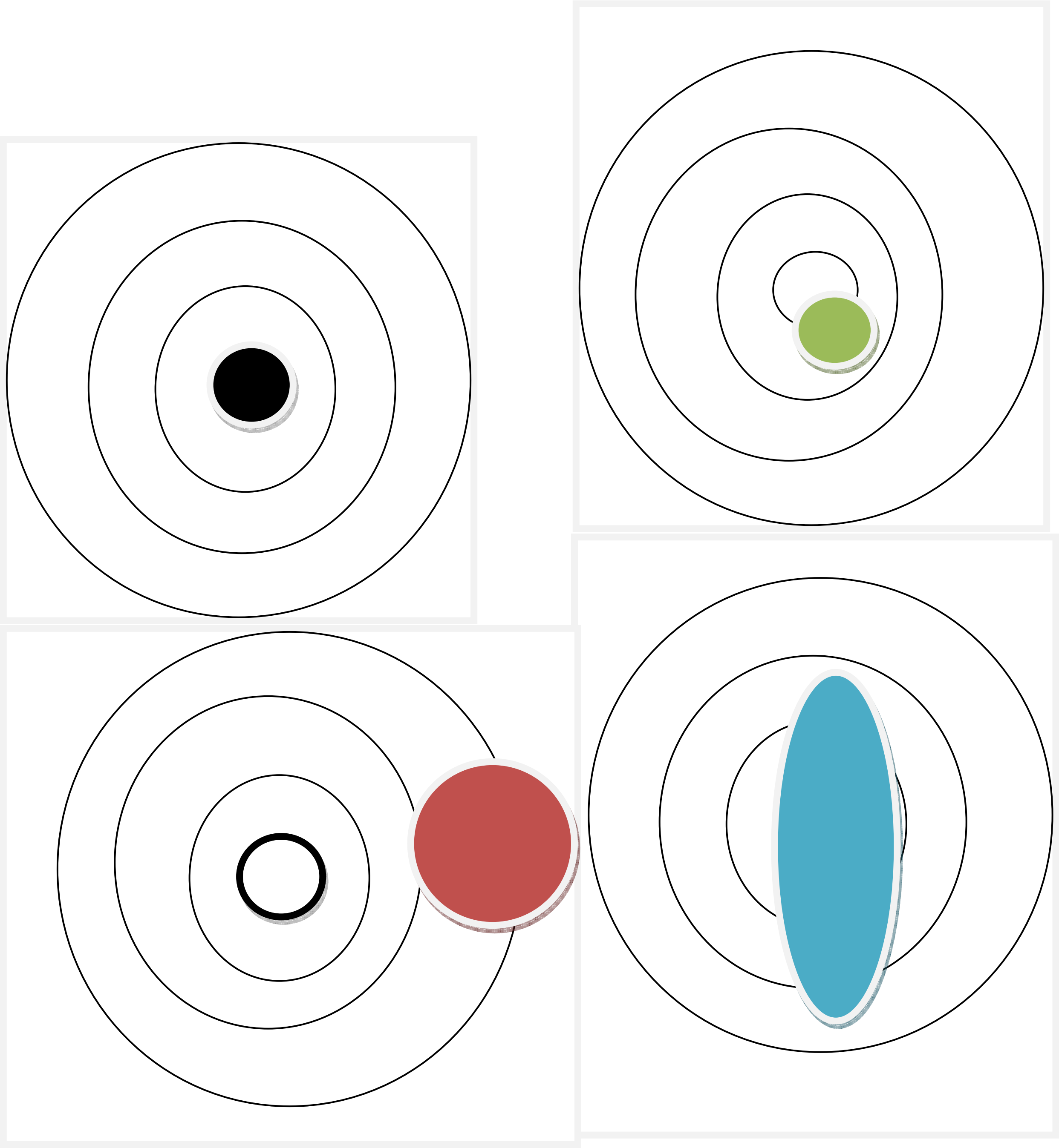
How would you get rid of bias?

Bias versus Precision:

Which do we want in a study, no bias or high precision?

Before we go into the BASIC principles of Design lets think of Bias and Precision based on the following example:

If the bull’s eye is our target, which scheme would you pick? Which have issues with bias and with precision?



This leads to the three basic tenets of DESIGN of EXPERIMENTS:

Randomization:

The procedure of selecting units at random from available units or assigning units to treatments at random. This reduces bias.

Replication:

Using more than one unit for a treatment for comparison. This establishes experimental error and also reduces bias.

Local Control:

Process of stratifying the units to homogenous groups or blocks and assigning treatments at random within the homogenous group. This reduces bias and reduces experimental error.

In ANY Design context think of these three tenets.

Lets go back to our original experiment:

Obviously picking 3 students from 1st year and assigning them a specific method is not the best way to go.

So we could do it as follows:

Recruit 9 students ***with roughly the same capabilities*** and assign 3 randomly to DT, randomly assign three to NN and three to BL. This way we deal with bias (through randomization) and deal with precision (by having more than one student assigned to each task – replication). But as all the students were “alike” or “homogenous” we do not have to worry about local control. THIS is an example of the Completely Randomized Design (CRD).

But now let us consider that 9 students who are in the same year with same capabilities is hard to find and best we can find is 3 from year 1, 3 from year 2 and 3 from year 3. Then our thought process is the three from a particular year have similar capabilities but are different from the ones from the other years. So here we use local control and assign the three from year 1 at random to Decision Trees, Neural Networks and Bayesian learning. So we use the difference among the years to control for the variability and randomize within each year. This type of experiment is called a BLOCK design and here the year is a block.

Some Definitions and Vocabulary in the context of DESIGN:

1. Factor, Levels, Treatment:

Factor: any substance or item whose effect on the data is to be studied. An experiment involving two or more treatment is called **factorial experiment**.

Our experiment had 1 factor.

Levels: values of the factor used in the experiment. The **levels of a factor** are the specific types or amounts of the factor that will actually be used in the experiment.

Our experiment had three levels (NN, DT, BL)

2. UNIT:

Experimental Unit: the unit to which the treatment is applied.

Here each student is a unit.

Observational unit (or Measurement unit): the unit on which the response is measured. In some cases, the observational unit may be different from the experimental unit - be careful!

The confusion between experimental and measurement units is a problem in field sciences but less in your fields.

3. Block: A homogenous group of units is a block.

   The student year is the block

4. Replicate: The multiple units used in the experiment is the replicate.

5. **Response Variable**: what outcome is being measured.

   Accuracy score on the task is our response

6. **Experimental error** is the variation among identically treated experimental units.

Model and Hypothesis:

Our Model in this framework using CRD is:

$Y_{ij} = \mu + \theta_i + \varepsilon_{ij}$

Here:

μ: is the overall effect on the accuracy

θi is the effect over and above the overall effect for each treatment.

ε_{ij}: is our random error component

What are we trying to test?

We are interested in seeing if there is a difference among the 3 treatments in terms of accuracy.

In Statistics we write this up as a hypothesis and test the hypothesis:

H0 : (NULL NYPOTHESIS) $\theta1=\theta2=\theta3=0$

HA: there is some difference among them.

Let us discuss Type I error and Power in the context of the hypothesis.

In computer sciences especially in determining the accuracy of a classifier the TRUTH is known. In Statistics hypothesis testing our 2 by 2 classification (similar to the confusion matrix is given as follows)

TRUTH (unknown) → Our Conclusion ↓	H0 (NULL) is TRUE	Research Hypothesis HA is TRUE
Reject H0	Type I error	
Do not reject H0		Type II Error

Type 1: error rejecting the null when null is true

Type II: Failing to reject null when the alternative is true.

In our context:

TYPE I: saying that there is a difference in mean accuracy among the three methods when in reality there is no difference among the methods.

Type I error is also called FALSE POSITIVES

Probability of Type I error is the False Positive rate

Type II: saying there is no difference among the three methods when there is a difference.

Power: 1-Prob(Type 1 error)

Sometimes this is called the TRUE POSITIVE RATE.

However, in Statistics since we do not KNOW what the truth is we can never really know what our TRUE type I error or Type II error is. What we can do is estimate the Probability of making a Type I error and Type II error.

Type I error and Type II error go hand in hand and cannot reduce them together. So what we do is ***fix*** our probability of Type I error and ***maximize*** power for that level of Type I error to fine the “most powerful test” in a given situation. **The idea is similar to a ROC plot of Specificity (TPR) against 1-sensitivity (FPR).** It is a tradeoff of having one versus the other. We want the best of both worlds in terms of a “good testing scheme”.

START THURSDAY

Consider the Basics of Testing in this case by putting in some numbers: In Table 1 and Table 2 I have some made up numbers that are realizations of the experiment we talked about. 5 students each in the three groups DT, NN, BL

Table 1

DT	NN	BL
5.90	5.51	5.01
5.92	5.50	5.00
5.91	5.50	4.99
5.89	5.49	4.98
5.88	5.50	5.02
5.90	5.50	5.00

Table 2

DT	NN	BL
5.90	6.31	4.52
4.42	3.54	6.93
7.51	4.73	4.48
7.89	7.20	5.55
3.78	5.72	3.52
5.90	5.50	5.00

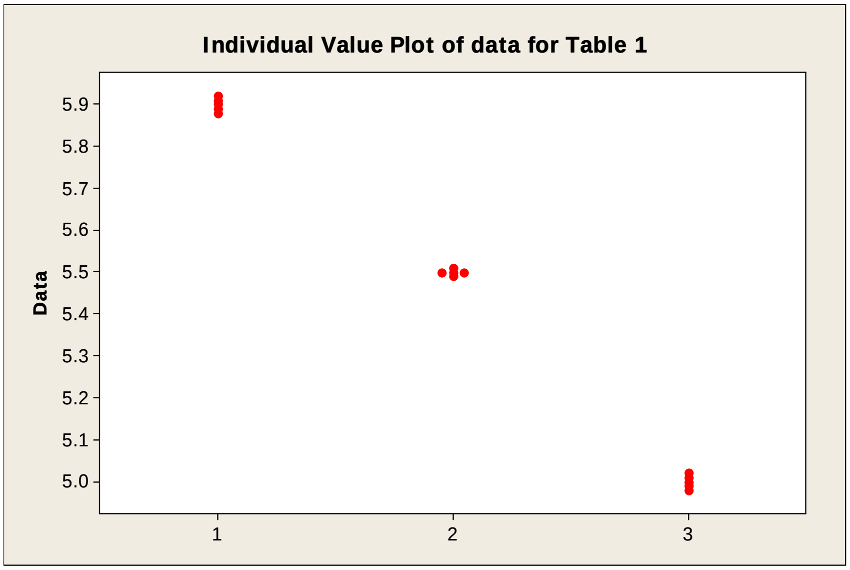
Just looking at the numbers can you say if the treatments are different in Table 1 and Table 2?

What allows you to say that they are different?

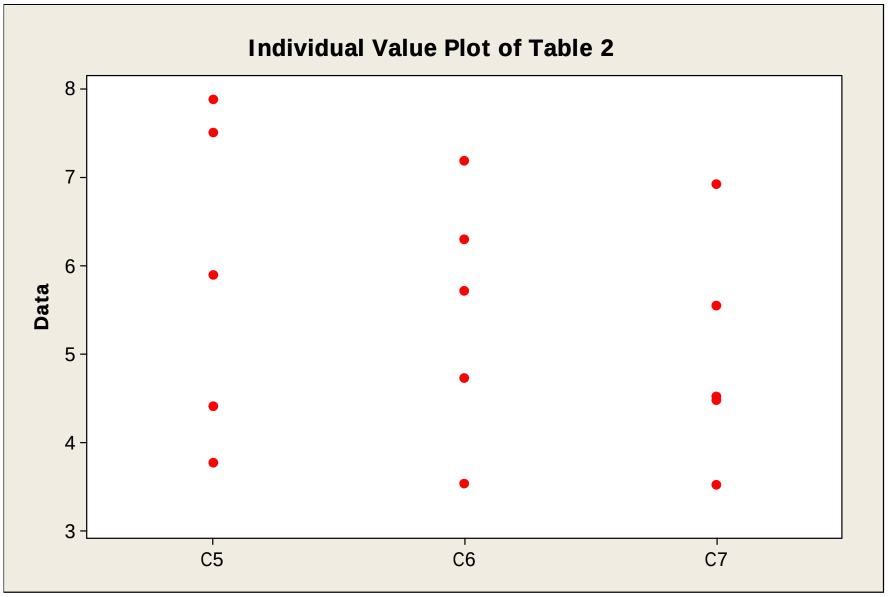
The answer is how the variability within is compared to variability ACROSS groups.

That’s exactly how we determine if the groups are different or not. We calculate the variability across the treatments and compare it with variability within. If the ratio is BIG we say that the treatments are different. This is PREMISE of the F test in ANOVA.

Lets look at the dot-plot of the data



Mean:	5.9	5.5	5.0
SD:	.016	.007	.016



Mean:	5.9	5.5	5.0
SD:	1.82	1.42	1.30

So in ANOVA what we do is find the WITHIN sample (sometimes called treatments) variance and compare that to the ACROSS sample variance. If the variation across the samples is MUCH greater than the variation within the samples, we can conclude that the samples means are different from each other. If the across sample variation is not big compared to the within sample variance, we cannot conclude that the means are different.

So the main work is to calculate the WITHIN and ACROSS sample variation. In ANOVA we call WITHIN variation the ERROR Variation and ACROSS variation the TREATMENT variance.

Measuring Variability within and across:

So how do we measure variability within a sample?

We pool it across all the samples.

So for the data in Table 1:

Lets first define some notation:

Level(i)	n_i	$\bar{y}_{i.}$	s_i
1	5	5.9000	0.0158
2	5	5.5000	0.0071
3	5	5.0000	0.0158

Here t=# of treatments=3

$N=n_1+n_2+n_3=15$

So the Variation within we use the notation s_w is:

$$S_W^2 = (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + (n_3 - 1)s_3^2$$

Pooling the variances across the three groups.

Now, how do we get Variability across samples?

We find the variances for the 3 sample means.

$$s_B{}^2 = (\bar{y}_{1.} - \bar{y}_{..})^2 + (\bar{y}_{2.} - \bar{y}_{..})^2 + (\bar{y}_{3.} - \bar{y}_{..})^2$$

Where, $\bar{y}_{..} = \frac{n_1\bar{y}_{1.} + n_2\bar{y}_{2.} + n_3\bar{y}_{3.}}{n_1+n_2+n_3}$

$N=n_1+n_2+n_3$

Lets calculate the terms for our example:

$s_W{}^2 = (5-1)(.0158)^2 + (5-1)(.0071)^2 + (5-1)(.0158)^2$=0.002200

$$\bar{y}_{..} = \frac{5(5.9)+5(5.5)+5(5.0)}{5+5+5} = 5.47$$

$$s_B{}^2 = (5.9-5.47)^2 + (5.5-5.47)^2 + (5.0-5.47)^2 = 2.033$$

So a logical choice for testing is the ratio $\frac{s_B{}^2/(t-1)}{s_W{}^2/(N-t)}$. If there are no treatment effects this ratio will be around 1 (variation across is about the same as variation between). If the ratio is MUCH greater than 1, we will inclined to think variation across is MORE than variation within.

In our case this is =5544.5 which is (a bit) bigger than 1? In general how big does this ratio have to be before we consider it too big (enough evidence, beyond reasonable doubt).

To find the cut-off we need to understand a BIT of distribution theory.

Distributions:

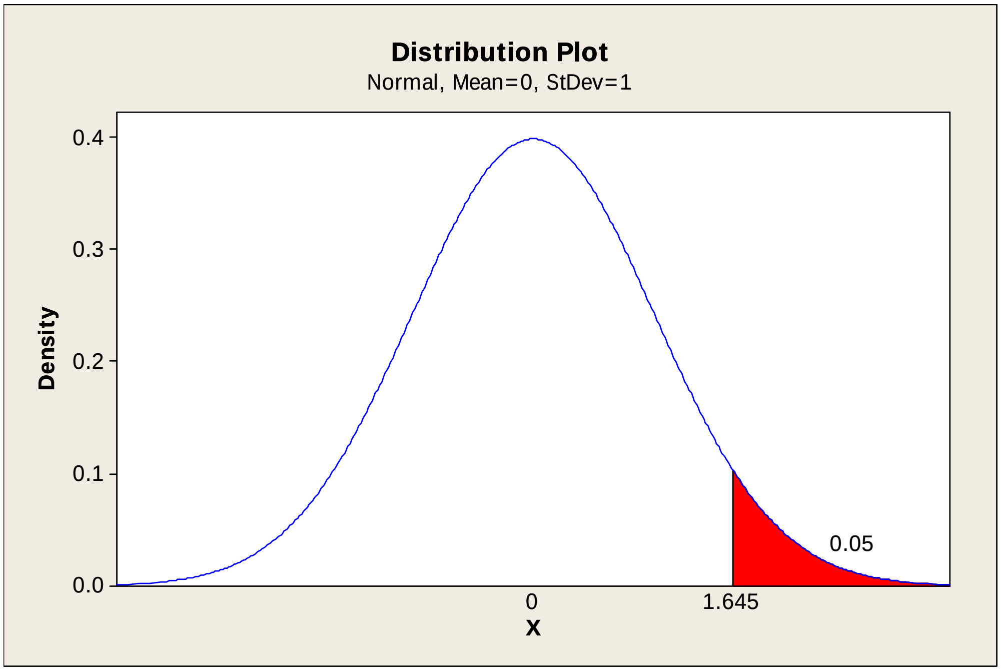
Normal (Gaussian or Laplace)

Let Y follow a Normal distribution with mean μ and standard deviation σ,

f(y) = $\dfrac{1}{\sigma\sqrt{2}}\exp(-\dfrac{1}{2}(\dfrac{y-\mu}{\sigma})^2)$

If Y* = a + bY, Y* follows a Normal distribution with E(Y*) = a + bμ and Var(Y*) = $b^2\sigma$.

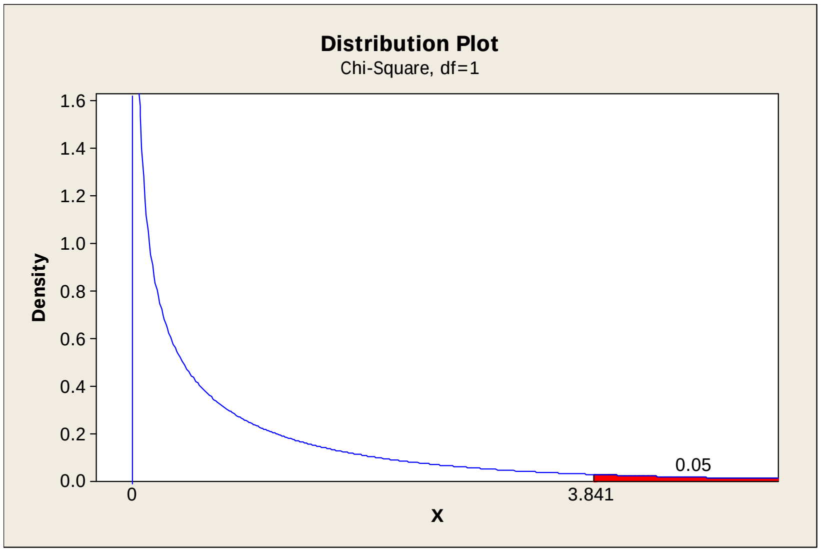
A normal distribution with mean 0 and variance 1 is called the STANDARD normal distribution.



Chi-square distribution:

Take a single Normal random variable Z_1 ~N(0,1).

Take the square of that, Z_1^2 what would it look like.



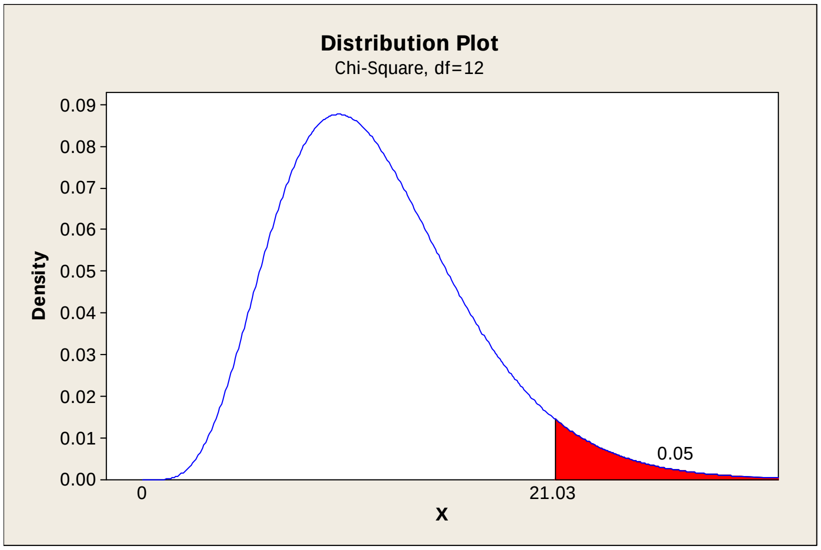
This is called a (central) chi-square with 1 degree of freedom.

Now, if we have Y ~N(μ, σ^2), then Y^2 follows a NON-central chi-square with 1 degree of freedom with a non-centrality parameter, λ = $\frac{\mu^2}{\sigma^2}$.

Let $z_1, \ldots, z_n$ be iid N(0,1). Then $X^2 = z_1^2 + \ldots + z_n^2$ follows a chi-square distribution with n degrees of freedom.

If yi ~ N(μ_i, σ^2), then $X^2 = y_1^2 + \ldots + y_n^2$ follows a NON chi-square distribution with n degrees of freedom and non-centrality parameter, λ = $\frac{\sum \mu_i^2}{\sigma^2}$.

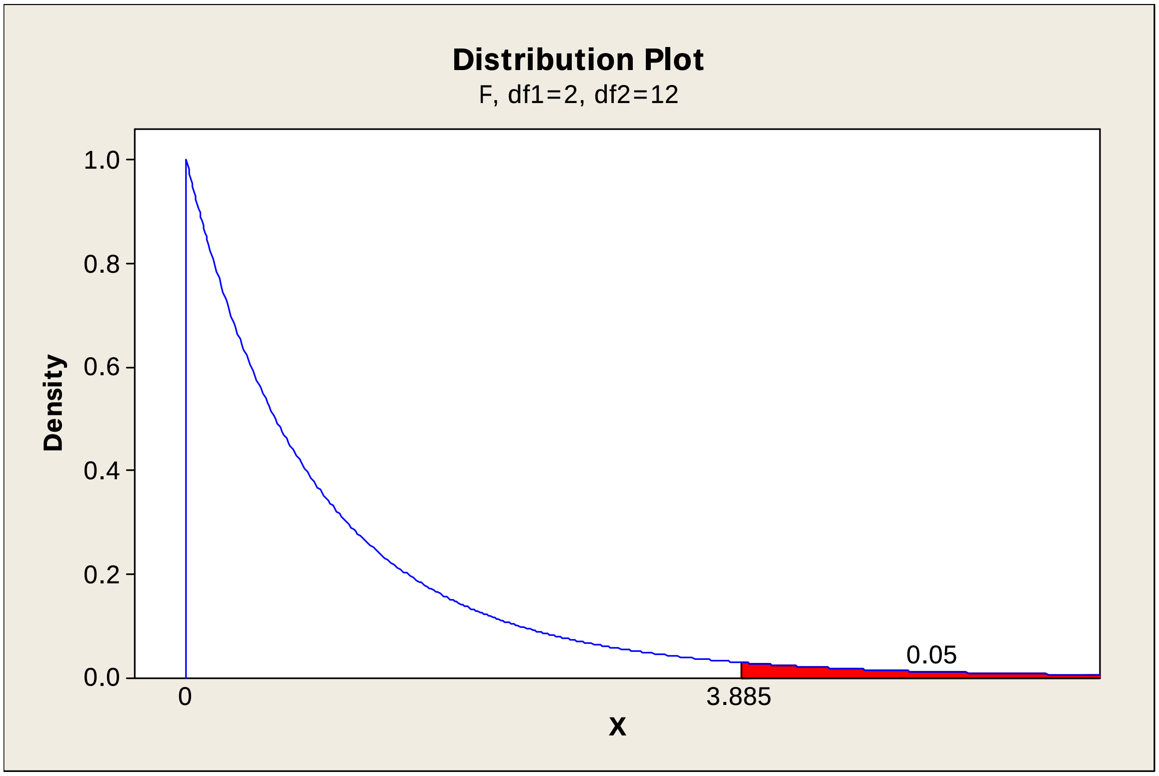
If there are some linear constraints upon the z's like: $z_1+\ldots.+z_n = a$ (some constant) then the degrees of freedom DECREASE by the number of constraints in the model.



F-distribution:

Let $\chi_1^2(\nu_1)$ and $\chi_2^2(\nu_2)$ follow INDEPENDENT (central) chi-square distributions with ν_1 and ν_2 degrees of freedom then:

$F = \dfrac{\chi_1^{\ 2}(\nu_1)/\nu_1}{\chi_2^{\ 2}(\nu_2)/\nu_2}$ ~ F distribution with ν_1 and ν_2 degrees of freedom.



Let $\chi_1^2(\nu_1,\lambda)$ and $\chi_2^2(\nu_2)$ follow INDEPENDENT chi-square distributions with ν_1 and ν_2 degrees of freedom then:

$F = \dfrac{\chi_1^{\ 2}(\nu_1)/\nu_1}{\chi_2^{\ 2}(\nu_2)/\nu_2}$ follow a F distribution with ν_1 and ν_2 degrees of freedom and non-centrality parameter λ.

So Back to our Problem:

We want to figure out how large is large for the ratio of $\frac{s_B{}^2/(t-1)}{s_W{}^2/(N-t)}$.

So we need to look at the model of ANOVA and assumptions to understand the distribution of the statistic.

Our assumption is that e_{ij} are independent and follows a Normal distribution with mean 0 and variance s^2.

Lets consider the following:

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

So,

$$\sum_{i,j}(y_{ij} - \bar{y}_{..})^2 = \sum_{i,j}(y_{ij} - \bar{y}_{i.})^2 + \sum_{j} n_i(\bar{y}_{i.} - \bar{y}_{..})^2$$

Total Sum Squares=Within Sum Squares + Between Sum Squares

Now let us reason this out,

Y is normal so $(y_{ij} - \bar{y}_{..})$ is also normal so $\frac{(y_{ij}-\bar{y}_{..})^2}{\sigma^2}$ is chi-square with $(N-1)$ degrees of freedom. Similarly $\frac{\sum_{i,j}(y_{ij}-\bar{y}_{i.})^2}{\sigma^2} = s_w^2$ is a chi-square with $(N-t)$ degrees of freedom and

$\frac{\sum_j n_i(\bar{y}_{i.}-\bar{y}_{..})^2}{\sigma^2} = s_B^2$ is a CENTRAL chi-square with $(t-1)$ degrees of freedom ***under the null hypothesis***.

So the ratio $\frac{s_B{}^2/(t-1)}{s_W{}^2/(N-t)}$ follows F with ((t-1), $(N-t)$) degrees of freedom under the NULL and follows F with (t-1,N-t) and $\lambda = \frac{\sum_{i=1}^{t} n_i(\theta_i)^2}{\sigma^2}$ under the alternative. So we reject the null when the λ is high.

In our example: n=15, t=3 so we are looking at F with 2 and 12 degrees of freedom. Which we saw from the graph was 3.88. So we reject the null hypothesis in Table 1.

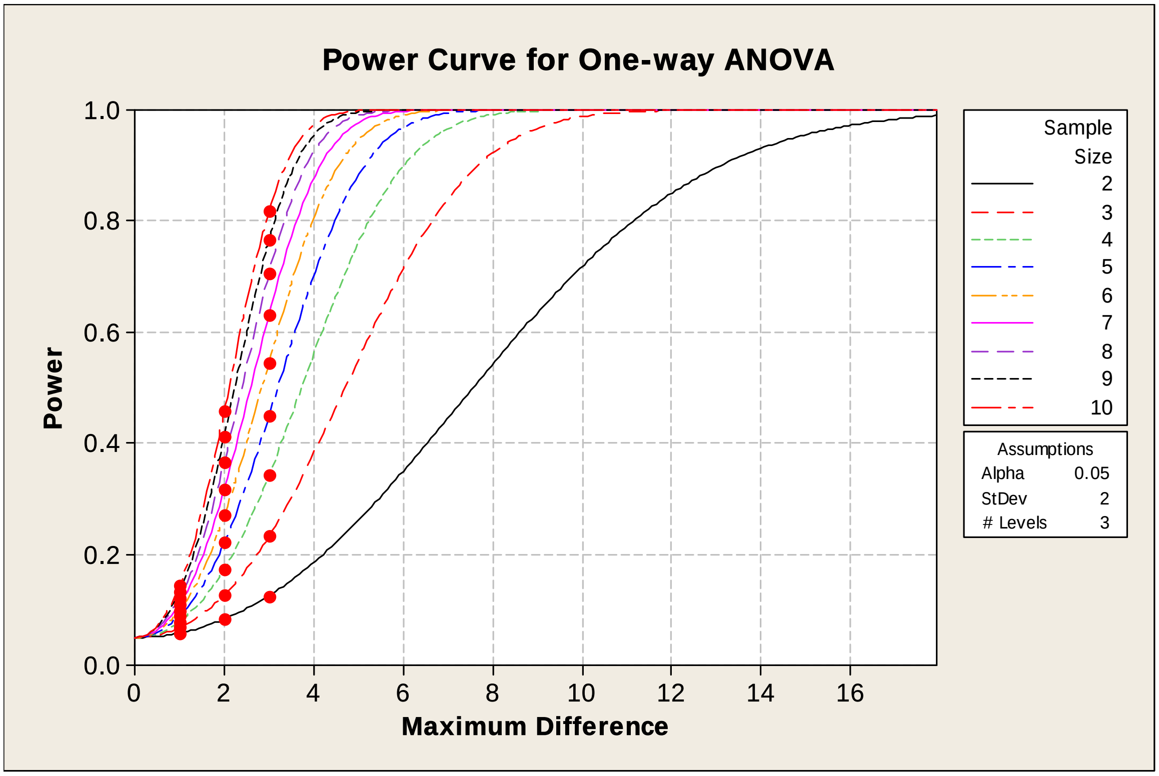
For Table 2 if you calculate F =.44 and then you would fail to reject the null.

Now let us think about what were the contributing reasons as to WHY we rejected the null hypothesis in this case. This leads us to the concept of power. Essentially power depends upon the non-centrality parameter and is a function of:

1. Total Sample size, N
2. Number of treatments, t
3. The difference among the true treatment means , $\sum_{i=1}^{t}(\theta_i^2)$
4. TRUE Standard deviation of the experiment, σ^2.

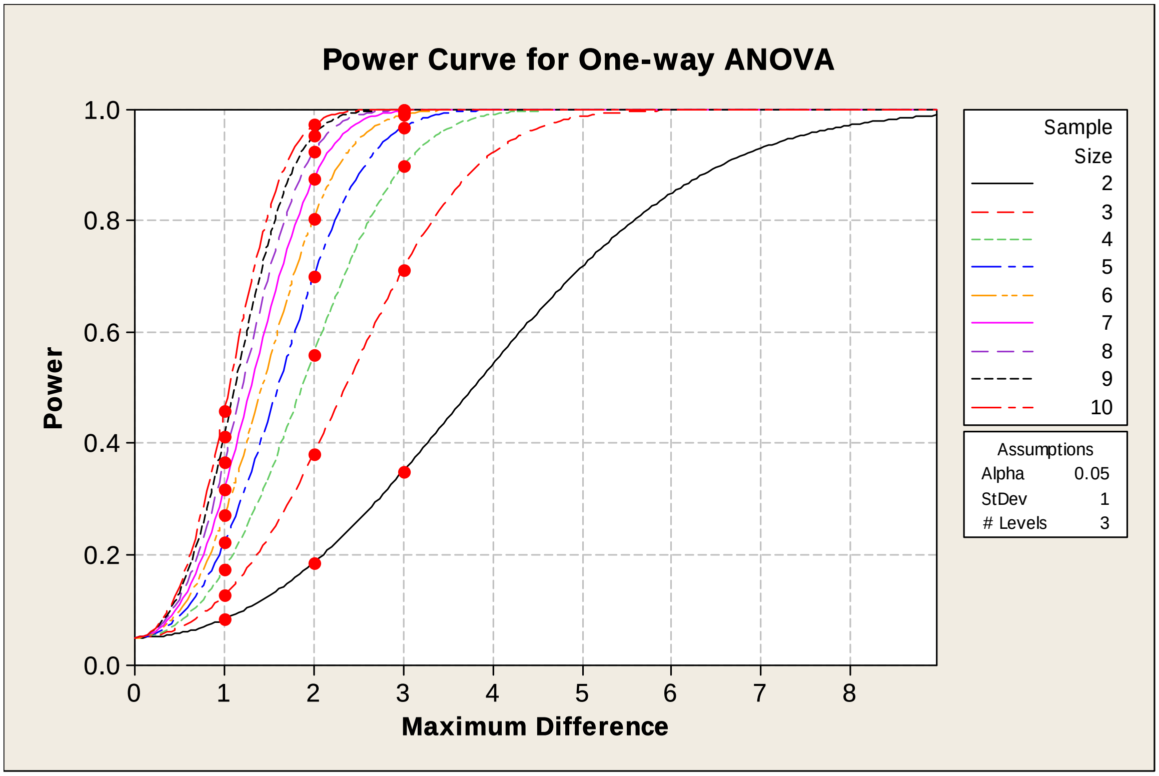
To determine sample size, for a given power we need to KNOW what the distance among the means are, how different each mean is from the others, what the standard deviation is and we can plot power as a function of sample size.

Power Curve with sigma=2



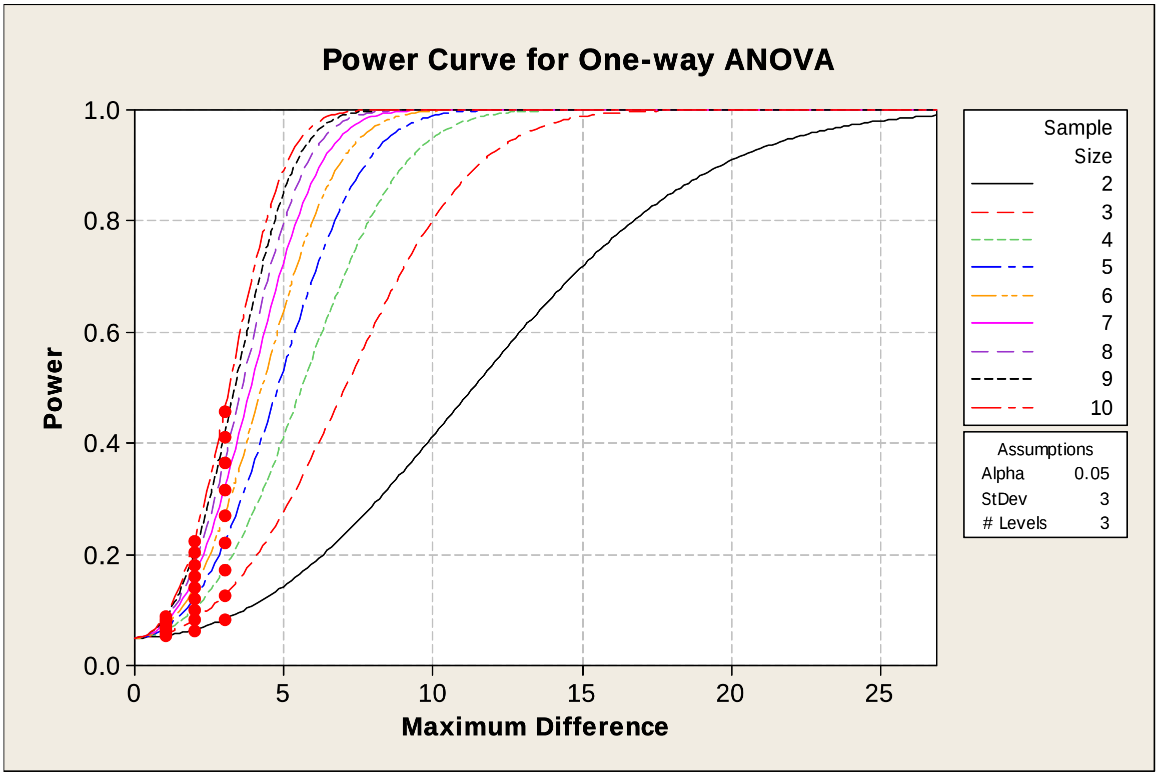
To have a Power >.8 for a maximum difference of 4 we would need n=6 observations per treatment

Power Curve with sigma=1



To have a Power >.8 for a maximum difference of 4 we would need n=3 observations per treatment

Power Curve with sigma=3



To have a Power >.8 for a maximum difference of 3 we would need n>10 observations per treatment.

So you see how important having some prior data is to be able to calculate sample size, based on power.