DOE Lecture 1

“What is An Experiment?”

Introduction

You are used to analyzing data and reporting a p-value; however, such analyses do not necessarily come from experiments. The data may be based upon:

1. Convenience Sample.
2. Observational Studies where the variable of interest may be "randomly dispersed in the population"
3. Probability Based Sampling methods.
4. Experimental designs. <- Causal Analysis

Why Experimental Design?

Mini-van Drivers

Average Speed:
64.2, 66.2, 67.0, 67.6,
67.8, 68.5, 68.8,
68.8, 68.9, 69.1

Sports Car Drivers

Average Speed:
69.2, 70.1, 70.2,70.7,
70.8, 71.0, 71.5
72.5, 73.5, 73.8

Question: Do people drive faster when they drive a sports car?

To compare two groups:

We can use a T-Test procedure

1. This is one of the standard analysis techniques along with ANOVA
2. Let $\alpha = 0.05$
3. Test for the differences in means the two groups
4. This is a standard HYPOTHESIS Test you have learned about previously.

Assumptions

1. Normal(ish) data
2. Variance
3. Random error IS independent and identically distributed

```
data car;
input car $4. speed;
cards;
van    64.22881
van    66.18087
van    66.98058
van    67.62792
van    67.8233
van    68.47134
van    68.77568
van    68.82076
van    68.93999
van    69.10898
scar   69.17223
scar   70.10115
scar   70.16689
scar   70.66927
scar   70.81493
scar   71.00318
scar   71.50214
scar   72.47056
scar   73.48653
scar   73.75543
;
```

In SAS this test is easy. Just load the data and run **proc ttest**.

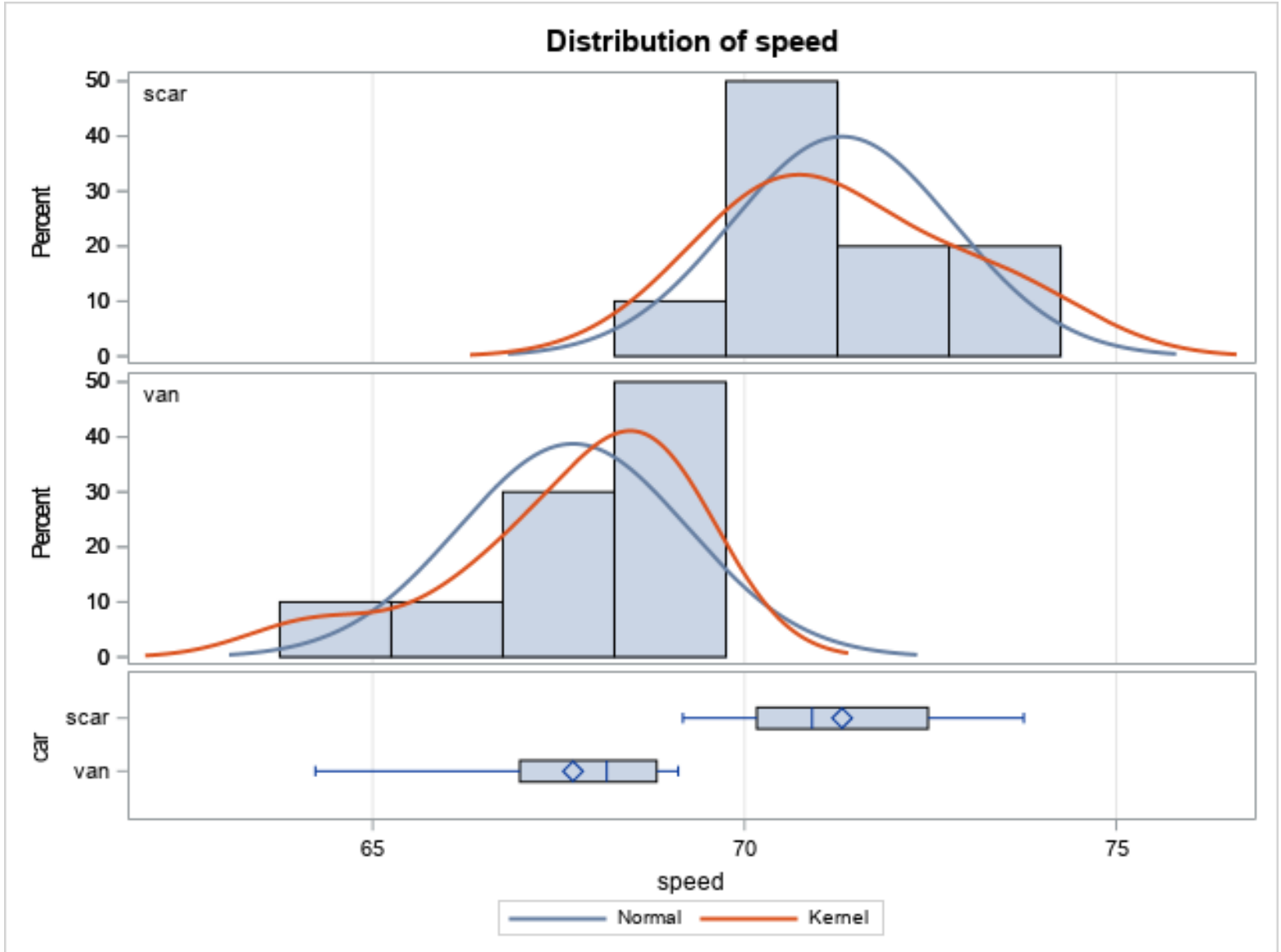
```
proc ttest data = car;
        class car;
        var speed;
run;
```

The TTEST Procedure Variable: speed

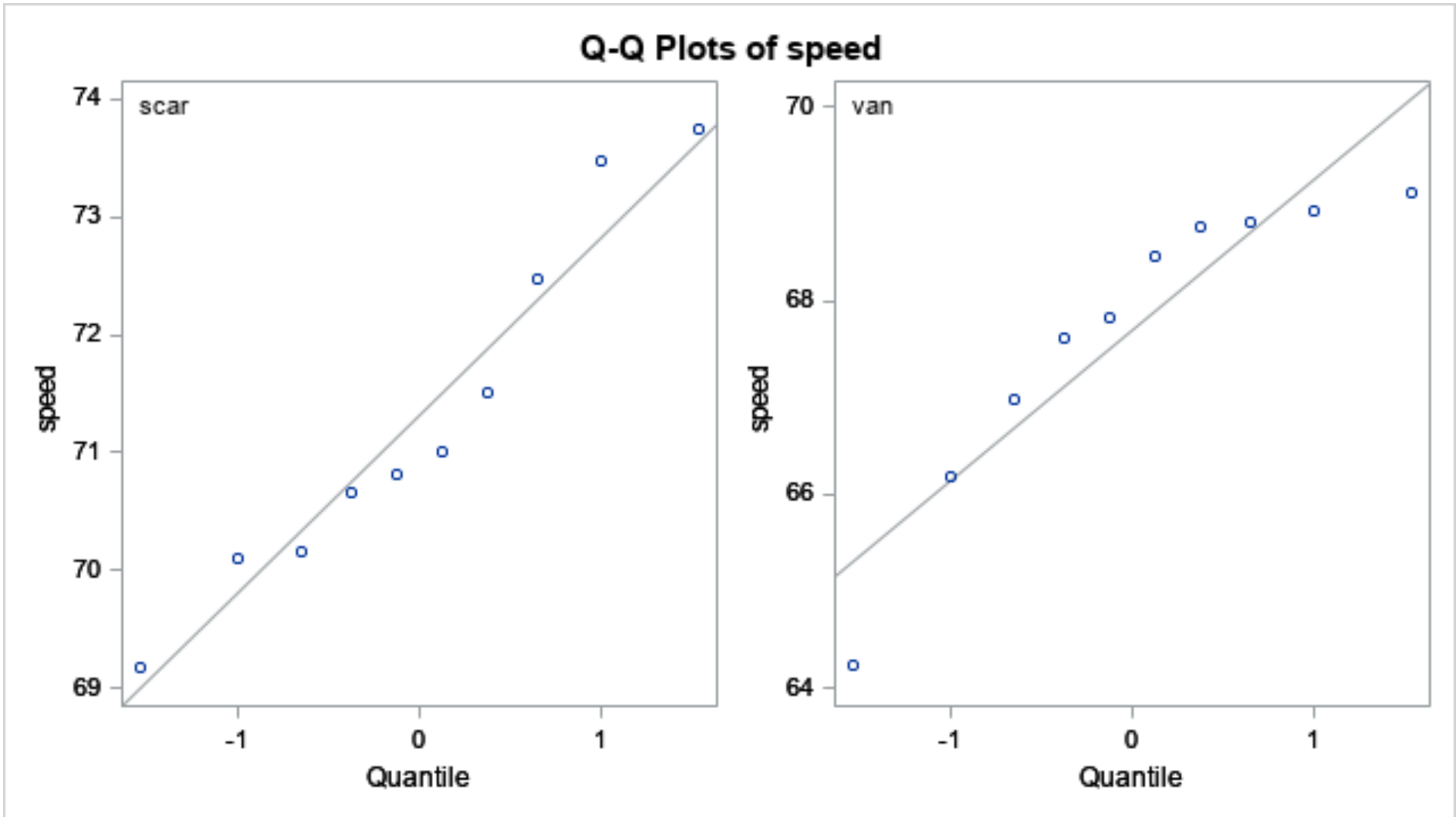
car	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
scar		10	71.3142	1.4977	0.4736	69.1722	73.7554
van		10	67.6958	1.5429	0.4879	64.2288	69.1090
Diff (1-2)	Pooled		3.6184	1.5205	0.6800		
Diff (1-2)	Satterthwaite		3.6184		0.6800		

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	18	5.32	<.0001
Satterthwaite	Unequal	17.984	5.32	<.0001

We can look at our assumptions:



Normality Assumptions:



After this observational analysis, we can conclude at the 95% level that when driving mini-vans people drive slower than when driving sports cars.

And, for this dataset, we could be dead wrong

WRONG!

1. This is an observational study, and I have no control over who is driving each car.
2. If it was a probability based method WE CAN'T conclude causality.
3. In this dataset, the car had nothing to do with it; rather, people who like driving fast drive sports cars.
4. What is worse we have no real way to tell we are wrong, all of our basic assumptions are reasonably met.
5. The problem is that I need to assign a driver to a car at random.
6. Random assignment of the "driver" in experimental design terminology "experimental unit" is done to minimize the effect of variables we do not know about.

From this perspective, each driver has a response to the treatment “the car” and we want to see if there is a difference between cars. **Randomization** is our attempt to account for uncontrolled variables that we do not know about.

Idea: I randomly assign the guy to the car.

Design of Experiments Basic Idea

1. Determine the main objectives:
    a) What is of interest to be studied? (e.g., differences car type in the example above)
    b) What is the measurement outcome. (e.g., speed in the example above.)
    c) Define the main comparisons. (e.g., Mini-van speed vs. Sports Car speed.)
    d) Define the main experimental unit. (e.g., driver)
2. Define the study plan.
    a) What is an important difference (e.g., is 0.1 mph important to detect or 10 mph).
    b) Determine study design
        i. Defines how variables that are to be controlled by the experimenter.
        ii. Figure out how big of a sample size you need to reasonably detect a difference. – Power.
    c) Optimality criterion. (discussed later)
3. Randomization – assign experimental units to “treatments.”

Design of Experiments Basic Idea

4. Analyze the data based upon the experimental design.
    a) Attempt to account for variables that were not controlled for.
    b) Check model assumptions.
    c) Only look at effects that were defined in the first step.

Nomenclature

Most of you will do A/B testing, but this is not what everyone will call it.

AB testing is also known as randomized controlled trial or a completely randomized design.

All of these things are the same thing!

The experimenter assigns “experimental units” to one treatment, which comes from two or more treatment conditions. The set of treatment conditions is called a “factor.” We want to see if there are differences between the treatments or groups of treatments, or if the factor in general changes the response.

Definitions:

Factor – Experimental variable of interest that potentially affects the response variable.

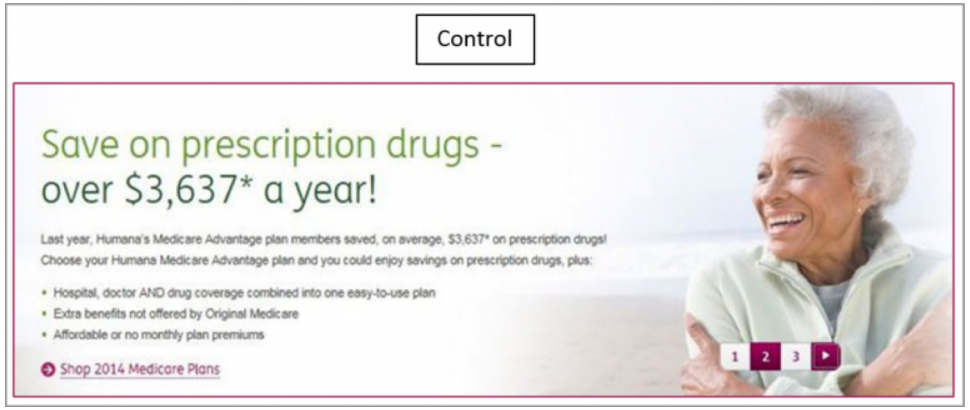
Treatment – Changes or levels of the factor. Usually only one treatment in a given factor is assigned to an experimental unit.

Experimental unit – The object to which the treatment is applied.

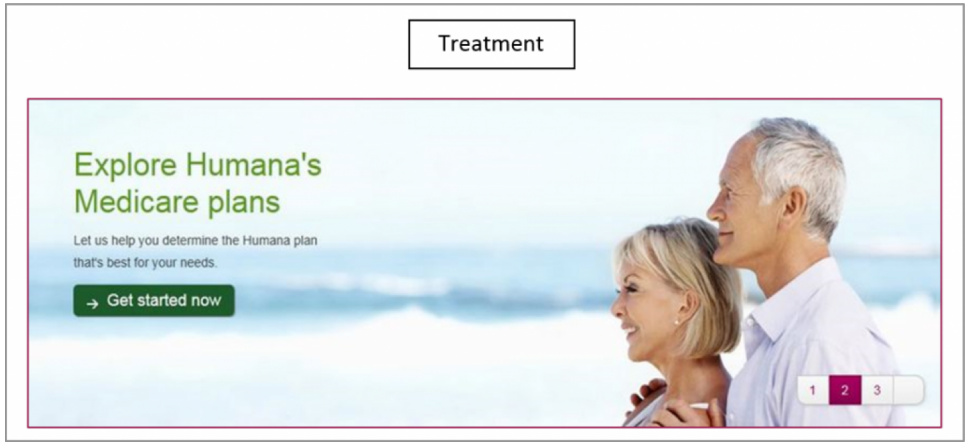
Blocks – Additional variables that should be accounted for, but can not be 'controlled for in the experiment.' These are nuisance variables that are not of interest, but could lead to erroneous results if not accounted for.

Example: Humana looked at changing its web page banner:

Web Visitors were randomly assigned to two different banners on Humana's website



Vs:



In this experiment, there is one **factor** to consider:

The changing banner on the website:

The factor had two **treatments**

1) The original banner
2) A newly redesigned banner.

Experimental Units:

Unique people visiting the website.

Randomization:

People are randomly assigned either the original banner or the newly redesigned banner.

Blocks:

None

Analysis Types

T-Test- Basic test when there is only one factor and two treatments.

ANOVA- Versatile method that incorporates all variability from the experiment. Multiple factors can be analyzed simultaneously and multiple treatments within a factor are possible. Additional covariates i.e blocks, can be accounted for.

Fishers Exact Test- Test used when for categorical data that are not normal and normal approximations above are not appropriate (usually small sample sizes).

Likelihood Ratio Tests: Used in situations were data are not normal to create ANOVA analogous tests.

When designing an experiment:

It is essential to first consider the type of data you are collecting.

This will allow you to :

1. Plan for the type of analysis
2. Estimate the power based upon the analysis type.
3. Determine the optimal sample size.

Other things:

Treatment Contrast- Planned tests between treatments in a factor. It can be between two treatments, or it can be between multiple treatments in a given factor. For example, if one is looking at three fertilizers, you can test to see if one fertilizer is better than both of the other fertilizers combined.

Power- Ability of one to detect a difference if it does exist. This is a function of – The sample size N, and the size of the actual difference.

Main effects- The effect directly attributable to the treatment.

Run – A full replicate of a given experimental design.

Estimate – The model based estimate of the treatment you are interested in.

Confounding or Aliased – Experimental design where you can not estimate a specific quantity.

Synergism or Higher order effect – Effect that changes with two factors.

Balanced Design – Special statistical term that implies the same number of observations are in each treatment group.

Factorial Design – Design where there are more than two factors. It usually is designed so that specific higher order effects (usually 2^{nd} order) are not aliased with other estimates. (more on this later).

Orthogonal Design – (more on this later)

Example 1: T-Test

A chair of an English department wanted to know the efficacy of two teachers in his program. Students were randomly assigned to a class and after the course a test was taken to determine their writing ability.

Factor: Teacher's teaching style

Treatment: Two college professors

Experimental Unit: College Freshman

Planned Comparison: Difference in scores using a paired T-Test

```
/*English data in class
  grp = 1 Teacher 1
  grp = 2 Teacher 2
*/
data class;
input class score;
cards;
1 35
2 52
1 51
2 87
1 66
2 76
1 42
2 62
1 37
2 81
1 46
2 71
1 60
2 55
1 55
2 67
1 53
;
```

```
/*t-test on the above example*/
proc ttest data = class;
        class class; /*class is the group
                       assignment 1 = Class 1*/
                        /*2 = Class 2*/
        var  score; /*variable for the t-test*/
run;
```

Proc T-Test Results

class	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
1		49.4444	41.4644	57.4245	10.3816	7.0123	19.8888
2		68.8750	58.5927	79.1573	12.2991	8.1318	25.0320
Diff (1-2)	Pooled	-19.4306	-31.1515	-7.7096	11.3169	8.3599	17.5151
Diff (1-2)	Satterthwaite	-19.4306	-31.3642	-7.4970			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	15	-3.53	0.0030
Satterthwaite	Unequal	13.823	-3.50	0.0036

Hypothesis H_0: Teachers are the same.

Hypothesis H_a: Teachers are different.

Conclude: At the $\alpha=0.05$, we conclude that the teachers have different effectiveness as evaluated by student's end of course writing proficiency.

More on Hypothesis testing:

For us: The key on hypothesis testing is to note we are trying to falsify the null hypothesis. That is we never accept the null as true we are just saying the null isn't true and there is more evidence for the alternative hypothesis.

Experimental design is thus: More about saying things are not true than saying something is true.

We want to control the probability that we decide the null isn't true, when it is true.

Decision Tree

State of Nature	Choose H_a	Choose H_o
H_0 True	Error – Type I α	1-α
H_a True	β	Error– Type II 1-β

On Power:

Power is the probably we reject the null when it is not true.

For normal models this value depends upon:

a) The size of the effect (usually mean difference in our case).
b) The variability of the population.
c) The type of test used and number of tests (multiplicity testing).
d) The sample size.

On Power:

For simple problems, including normal models SAS can calculate the power we reject the null for a series of situations.

This can help us in two ways:

1. Design a study with a high probability of success. (save money)
2. Find out the probability (after the fact) that our result is true. That is if we are not getting any differences, what is the probability given our sample size that we actually observed something.

proc power in SAS

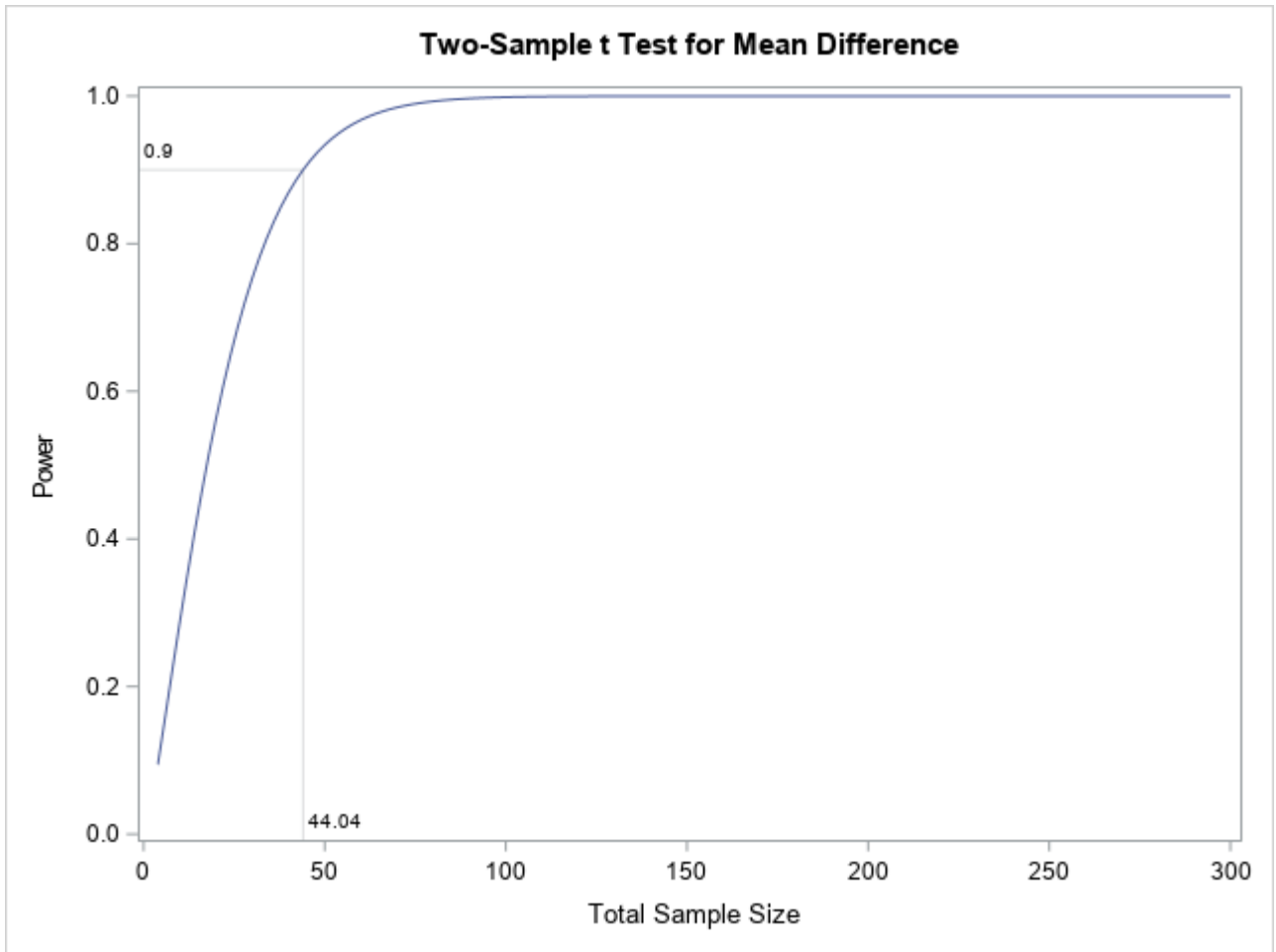
```
proc power;
   twosamplemeans
      test=diff
      meandiff = 5 to 25 by 5 /* vary the mean difference*/
      stddev = 5 to 15 by 5   /* vary the standard deviation of the groups*/
      ntotal = . /*find the n */
      power = 0.9;  /*we want a 90% probability to find a differences
                              if the above is true*/
run;
```

Computed N Total				
Index	Mean Diff	Std Dev	Actual Power	N Total
1	5	5	0.912	46
2	5	10	0.903	172
3	5	15	0.901	382
4	10	5	0.929	14
5	10	10	0.912	46
6	10	15	0.904	98
7	15	5	0.939	8
8	15	10	0.917	22
9	15	15	0.912	46
10	20	5	0.948	6
11	20	10	0.929	14
12	20	15	0.903	26
13	25	5	0.993	6
14	25	10	0.932	10
15	25	15	0.913	18

Our power was about 93%

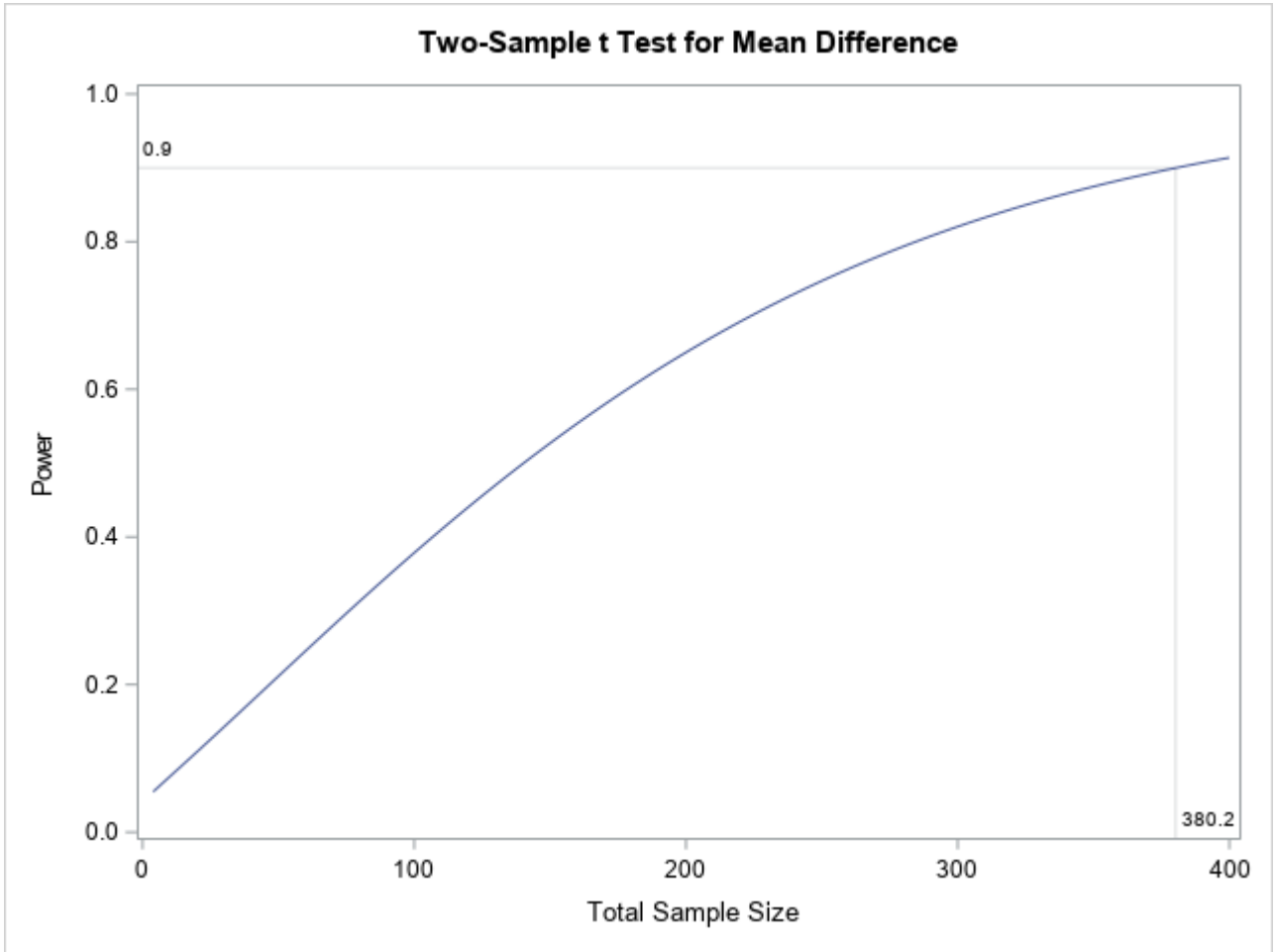
Power Curve

```
proc power;
      twosamplemeans
            test=diff
         meandiff = 5
         stddev = 5   /* vary the standard deviation of the groups*/
               NTOTAL = 1 to 300 by 5
         POWER = .;
               /*now look at the graph*/
               /* 90% is my cutoff ref=0.9)*/
              PLOT X=N MIN=1  MAX=300 STEP=1 MARKERS=none YOPTS=
(REF=0.9 CROSSREF=yes);
run;
```


Two-Sample t Test for Mean Difference
Power
1.0
0.8
0.6
0.4
0.2
0.0
0.9
44.04
0
50
100
150
200
250
300
Total Sample Size

Different SD

```
proc power;
      twosamplemeans
              test=diff
         meandiff = 5
         stddev = 15
        NTOTAL = 1 to 400 by 5
         POWER = .;
               /*now look at the graph*/
               /* 90% is my cutoff ref=0.9)*/
              PLOT X=N MIN=1  MAX=400 STEP=1 MARKERS=none YOPTS= (REF=0.9
CROSSREF=yes);
run;
```


Two-Sample t Test for Mean Difference
1.0
0.8
0.6
0.4
0.2
0.0
Power
0.9
380.2
0
100
200
300
400
Total Sample Size

Example 2:

A study was designed to determine the effect of copper concentrations in water on the lifetime of the Daphinia Magna. Fifteen Daphinia were randomized to three treatments: 0 µg/L , 20 µg/L, and 40 µg/L.

Factor: Copper Exposure

Treatments: 0 µg/L , 20 µg/L, and 40 µg/L.

Randomization: Fifteen Daphinia assigned to the treatments random assignment.

Example: Data

0 µg/L	20 µg/L	40 µg/L
60	58	40
90	74	58
74	50	25
82	65	30
	68	42

Example (Cont):

Contrasts:

1. Difference between 0 µg/L and 20 µg/L

Overall Test Level:

$\alpha = 0.05$

```
/* data for example 2
   in Lecture 1*/
data dubia_exp;
       input dose age;
       cards;
       0  60
       0  90
       0  74
       0  82
       20  58
       20  74
        20  50
       20  65
       20  68
;
```

```
proc ttest data = dubia_exp;
       class dose;
run;
quit;
```

proc ttest results

Our test was not significant at the $\alpha = 0.05$ level

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	7	1.84	0.1079
Satterthwaite	Unequal	5.3441	1.77	0.1330

dose	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
0		76.5000	56.1431	96.8569	12.7932	7.2472	47.7002
20		63.0000	51.4853	74.5147	9.2736	5.5561	26.6483
Diff (1-2)	Pooled	13.5000	-3.8246	30.8246	10.9218	7.2212	22.2288
Diff (1-2)	Satterthwaite	13.5000	-5.7232	32.7232			

There is not enough evidence to conclude a difference between between the 0 µg and the 20 µg doses? Did we have enough power?

Let’s see using proc power:

Assume:

1. The difference we saw was real.
2. The observed variance is real.
3. We are using a t-test for the difference

```
/*what was the (approximate) power to
detect a difference between
  the two groups. Note: our sample size
per group is not even*/
proc power;
      twosamplemeans
        test=diff
        meandiff = 13.5
        stddev = 10.9
        ntotal = 9
        power = .;
run;
```

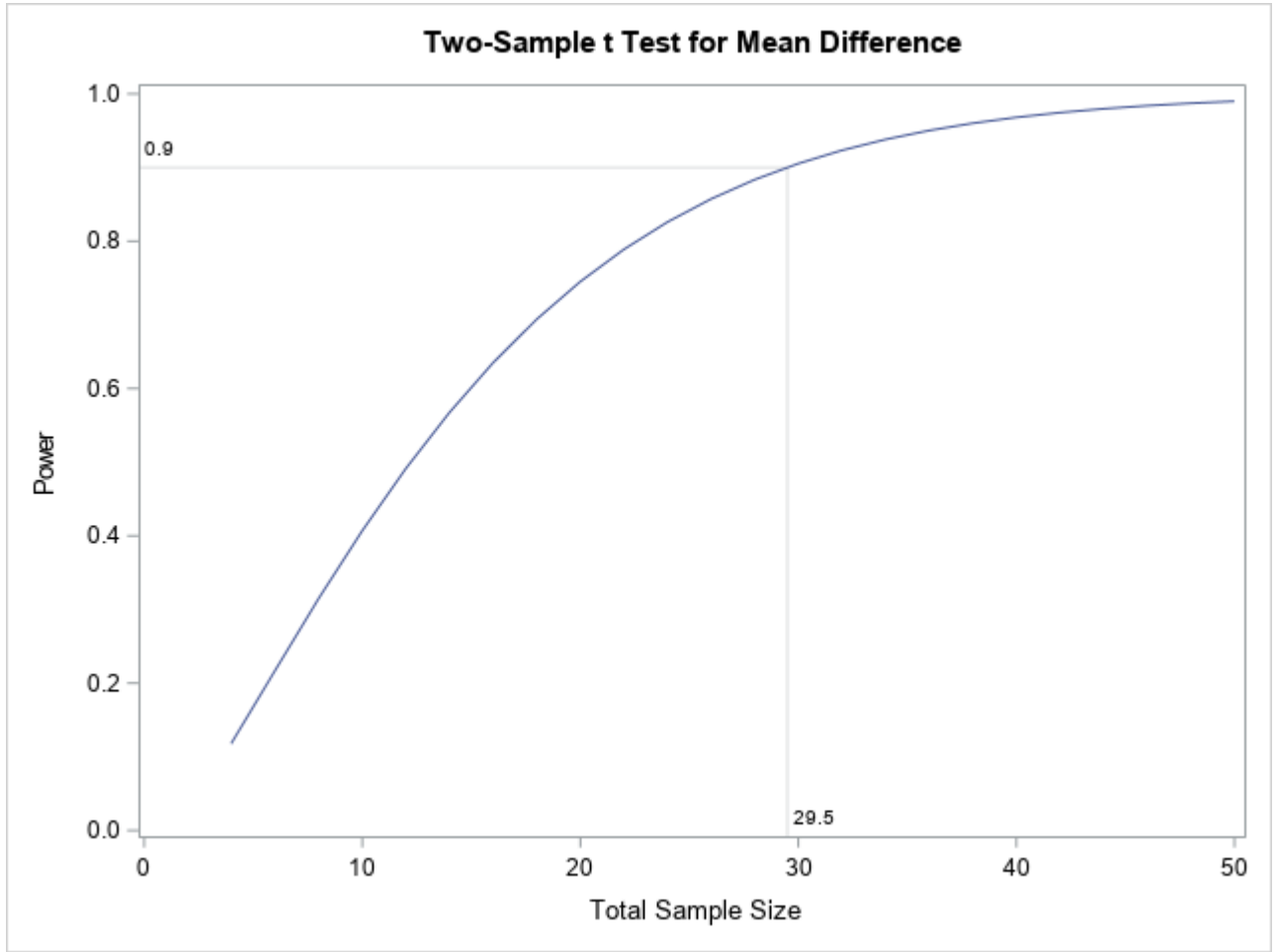
Fixed Scenario Elements	
Distribution	Normal
Method	Exact
Mean Difference	13.5
Standard Deviation	10.9
Nominal Total Sample Size	9
Actual Total Sample Size	8
Number of Sides	2
Null Difference	0
Alpha	0.05
Group 1 Weight	1
Group 2 Weight	1

We only had a 31.5% chance to detect a difference.

Computed Power
Power
0.315

What is the sample size we need for a 90% probability to detect the difference?

```
/* Again assuming the mean is true, look at this over
a range
of values. */
proc power;
      twosamplemeans
         test=diff
         meandiff = 13.5
         stddev = 10.9
         ntotal = 5 to 50 by 2
         power = .;
         PLOT X=N MIN=1  MAX=50 STEP=1 MARKERS=none
YOPTS= (REF=0.9 CROSSREF=yes);
run;
```


Two-Sample t Test for Mean Difference
Power
1.0
0.8
0.6
0.4
0.2
0.0
0.9
29.5
0
10
20
30
40
50
Total Sample Size

If this difference is true, our study was too **small**!

We really needed about 30 observations (15 per group) to detect this difference at a high probability.

If this is an important difference: **We wasted money**!