POWER

Matthew W. Wheeler

Power

If we are designing an experiment, we are spending money.

We want to have a reasonable probability (not assured) that we find something, if it exists.

power: The probability that we reject the null hypothesis in favor of our alternative.

Power depends upon:

1. The true state of nature, which in our case (for now) is the actual mean and the variance of the response. This is often called the **effect size**.
2. Our Type I error rate. If we are willing to make a type I error more frequently, our power increases.
3. Our sample size.
4. Our experimental design. Certain designs are "better." (More on that later)

When designing a study, we can only control 2-4. We have to guess at 1.

SAS has two procs that help us with this:

proc power

and

proc glmpower

For more complicated experiments, we need to simulate the power.

We are going to focus on proc power now, and we will look at the other two later.

From SAS help

Syntax: POWER Procedure

The following statements are available in PROC POWER:

PROC POWER <options> ; **<- invoking statement**
LOGISTIC <options> ;
MULTREG <options> ; **<- for regression analysis**
ONECORR <options> ;
ONESAMPLEFREQ <options> ; **<- simple test comparing proportions to a known quantity**
ONESAMPLEMEANS <options> ; **<- simple ttest comparing proportions to a known quantity**
ONEWAYANOVA <options> ;
PAIREDFREQ <options> ;
PAIREDMEANS <options> ;
PLOT <plot-options> </ graph-options> ;
TWOSAMPLEFREQ <options> ; **<- paired test for two proportions A/B testing conversions.**
TWOSAMPLEMEANS <options> ; **<-paired ttest for two unknown means.**
TWOSAMPLESURVIVAL <options> ;
TWOSAMPLEWILCOXON <options> ;

We are going to focus on some of the above options with the goal of figuring out the sample size required to achieve a reasonable power.

Reasonable power is usually between 80% and 90%.

For clinical research involving humans or animals, a power of 80% is usually considered ‘ethical.’

Example 1

A marketing campaign has been designed to increase customer spending. Currently customers spend an average of $200 a month with a standard deviation of $20, the campaign is considered successful if customers spend $25 more a month. We are going to design an experiment to see if the campaign should be rolled out nationwide. How many people should be recruited to see if this effect is found with 90% probability?

Our analysis is a simple t-test.

```
PROC POWER;
    ONESAMPLEMEANS TEST=t
    NULLMEAN=200
    MEAN=225
    STDDEV=20
    ALPHA=0.05
    POWER=0.9
    NTOTAL=.;
RUN;
```

Fixed Scenario Elements	
Distribution	Normal
Method	Exact
Null Mean	200
Alpha	0.05
Mean	225
Standard Deviation	20
Nominal Power	0.9
Number of Sides	2

Computed N Total	
Actual Power	N Total
0.906	9

Example 1 (continued)

Now assume we don’t know the standard deviation, it could be anywhere between 10 to 50.

```
STDDEV= 10 to 50 by 5
```

Computed N Total

Index	Std Dev	Actual Power	N Total
1	10	0.982	5
2	15	0.953	7
3	20	0.906	9
4	25	0.911	13
5	30	0.915	18
6	35	0.905	23
7	40	0.901	29
8	45	0.908	37
9	50	0.900	44

Example 1 (continued)

Now assume we can only afford to recruit 20 people. Do we have enough power?

Change

```
POWER=0.9 to POWER=.
and
NTOTAL=. to NTOTAL=20 ;
```

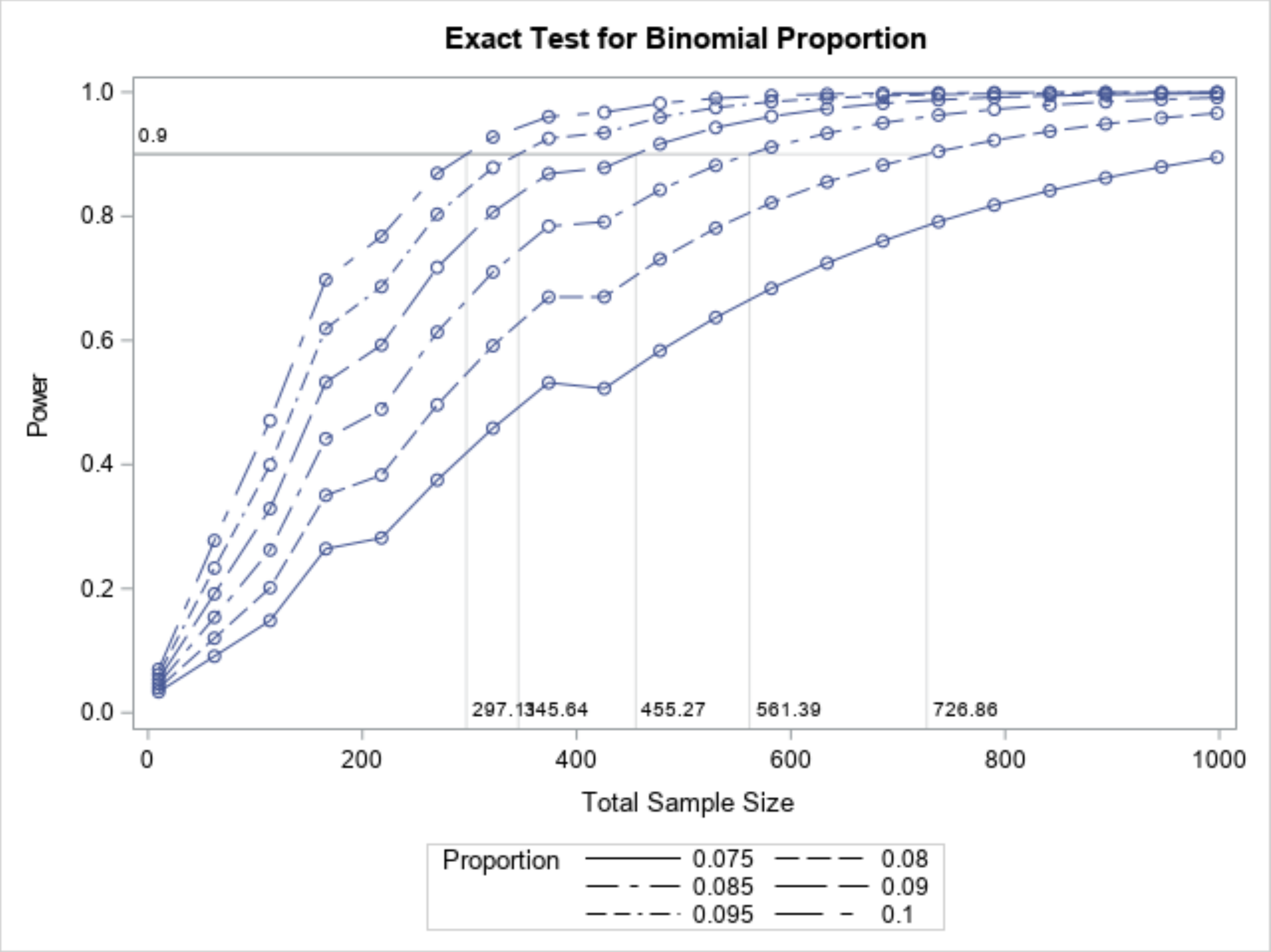
Computed Power		
Index	Std Dev	Power
1	10	>.999
2	15	>.999
3	20	>.999
4	25	0.989
5	30	0.942
6	35	0.858
7	40	0.755
8	45	0.655
9	50	0.565

Example 2

A web page site currently has a 5% conversions rate. The redesigned website is hoped to have a 7.5% conversion rate. What is the sample size required power is required to see if the new website has the desired effect?

Note: The data are proportion data, so we can do what is called an exact test.

```
PROC POWER;
      ONESAMPLEFREQ
      NULLP=0.05
      P=0.075 TO 0.1 BY 0.005 /*VARY THE CRITICAL EFFECT FROM
                                            0.075 TO 0.1*/
      ALPHA=0.05
      POWER= .
      NTOTAL= 900;
      PLOT X=N min=10 max=1000 yopts = (ref=0.90 crossref=yes)
                  ;*PLOT THE POWER FROM 10 TO 1000;
RUN;
```


Exact Test for Binomial Proportion
Power
1.0
0.8
0.6
0.4
0.2
0.0
0.9
297.13
345.64
455.27
561.39
726.86
0
200
400
600
800
1000
Total Sample Size
Proportion
0.075
0.08
0.085
0.09
0.095
0.1

The Jagged curve is because the test is no longer an approximation (i.e. normal approximation), but it is based upon a different distribution. This is known as a saw tooth power function.

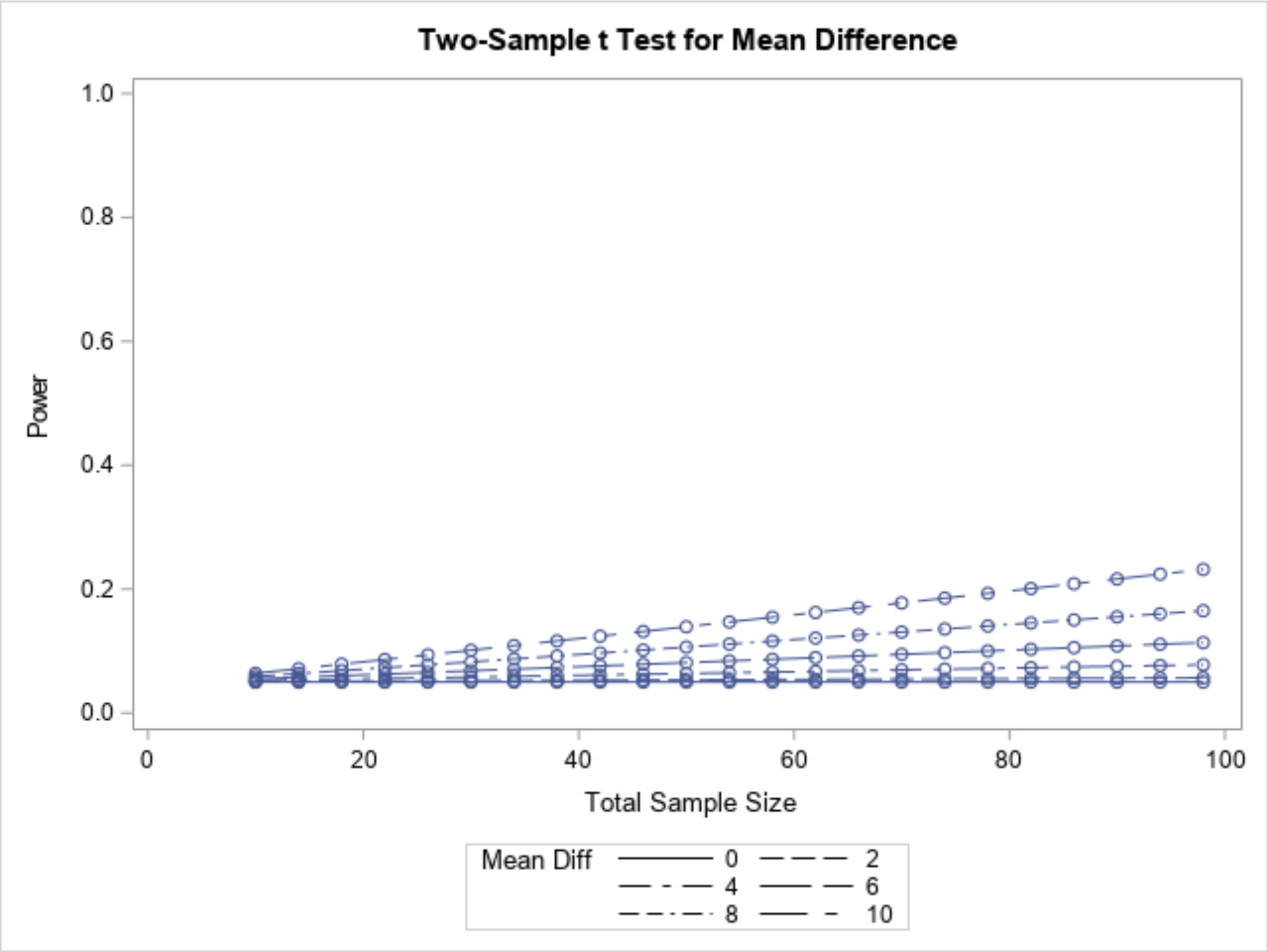
Until know we have assumed there is one unknown. Let’s assume there are two unknown means.

Example 3

A two different drugs are designed to lower cholesterol. We are interested in determining if the plasma total cholesterol is different between the two drugs. We know that the population serum total cholesterol is has a standard deviation of about 43 mg/dl in the general population.

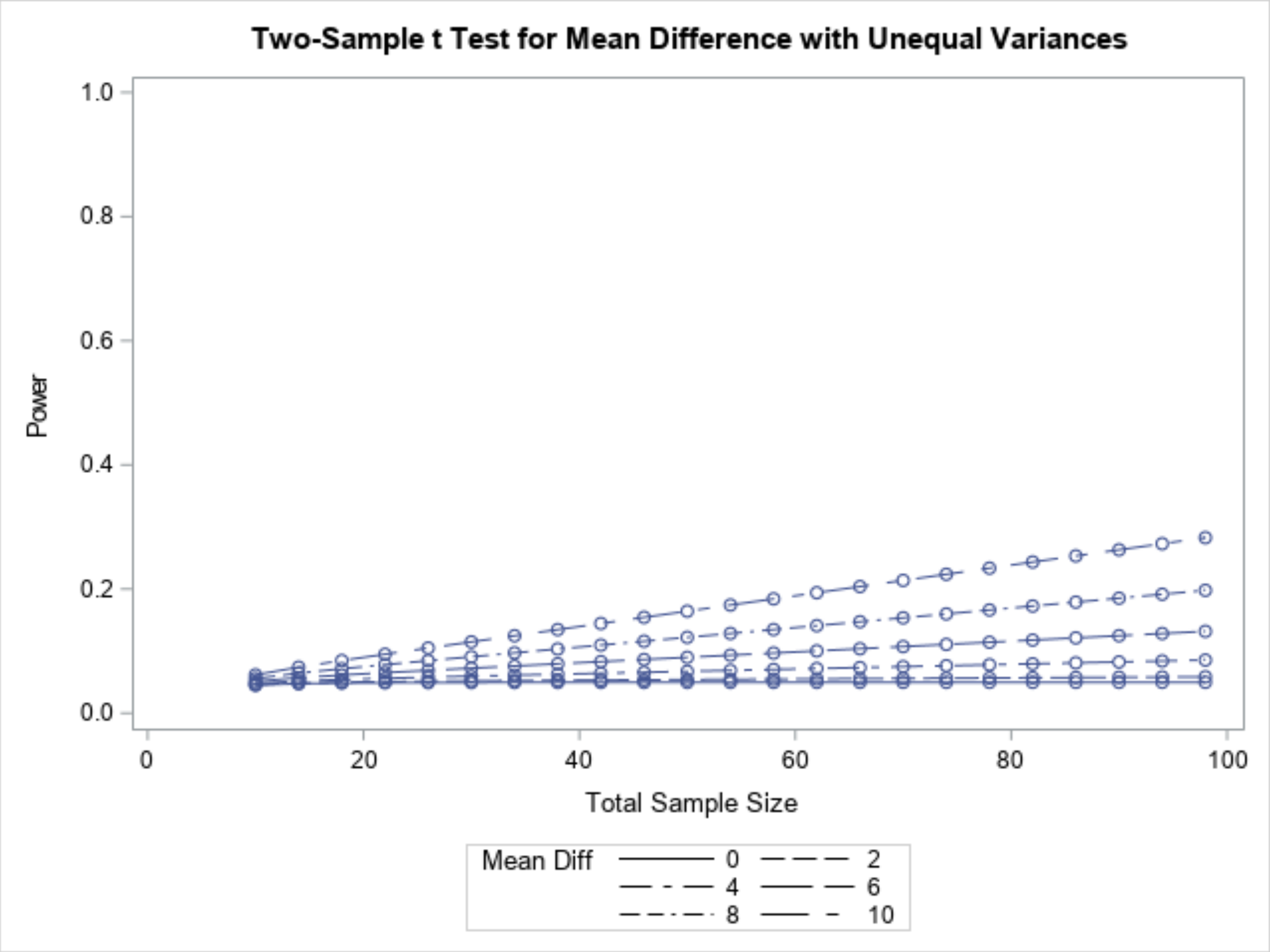
```
*EXAMPLE 1;
*WHAT WE REALLY CARE ABOUT IS A DIFFERENCE IN THE
MEAN
* WE DON'T HAVE TO SPECIFY EACH MEAN INDIVIDUALLY;
* QUESTION WHAT IS MY POWER WHEN I CHANGE N OVER THE
RANGE
* 20 TO 100;
PROC POWER;
       TWOSAMPLEMEANS
       MEANDIFF = 0 TO 10 BY 2
       STDDEV = 40
       ALPHA = 0.05
       POWER = .
       NTOTAL = 20 TO 100 BY 20;
       PLOT X=N min=10 max=100 yopts = (ref=0.90
crossref=yes);
RUN;
END;
```

My power up to a sample of 100 is pretty low.



What if one drug not only reduces total Cholesterol but also decreases the variability of the response?

```
*COMPARE THE NORMAL TEST ABOVE TO THE CASE
*WHEN THERE IS UNEQUAL VARIANCES
*;
PROC POWER;
       TWOSAMPLEMEANS  TEST=diff_satt
/*satterthwaite approximation unequal
variances*/
       MEANDIFF = 0 TO 10 BY 2
       GSTDS = (40,30)
       ALPHA = 0.05
       POWER = .
       NTOTAL = 20 TO 100 BY 20;
       PLOT X=N min=10 max=100 yopts =
(ref=0.90 crossref=yes);
RUN;
END;
```

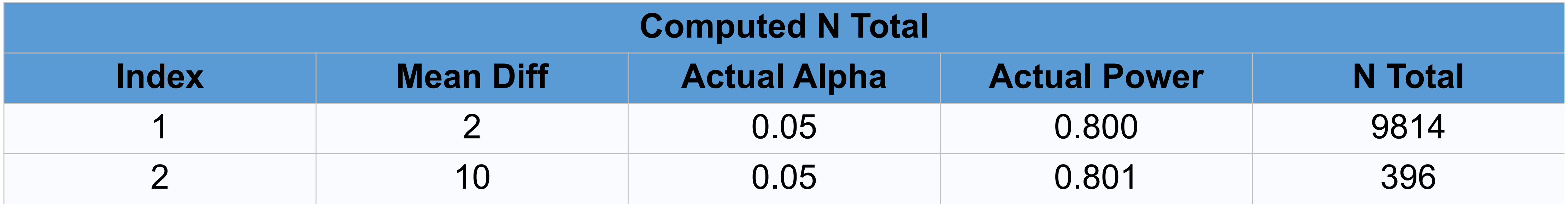

Two-Sample t Test for Mean Difference with Unequal Variances
Power
1.0
0.8
0.6
0.4
0.2
0.0
0
20
40
60
80
100
Total Sample Size
Mean Diff
0
2
4
6
8
10

We want 80% power

```
*WE DON'T HAVE A LOT OF POWER, WHAT IS OUR
SAMPLE
SIZE IF WE WANT 80% POWER?;
PROC POWER;
      TWOSAMPLEMEANS  TEST=diff_satt
/*Satterthwaite approximation unequal
variances*/
      MEANDIFF = 2 TO 10 BY 8
      GSTDS = (40,30)
      ALPHA = 0.05
      POWER = .80
      NTOTAL = .;
RUN;
END;
```

If we want to detect a difference of 2, we need a ton of people!

The question becomes is it worth it?



Computed N Total

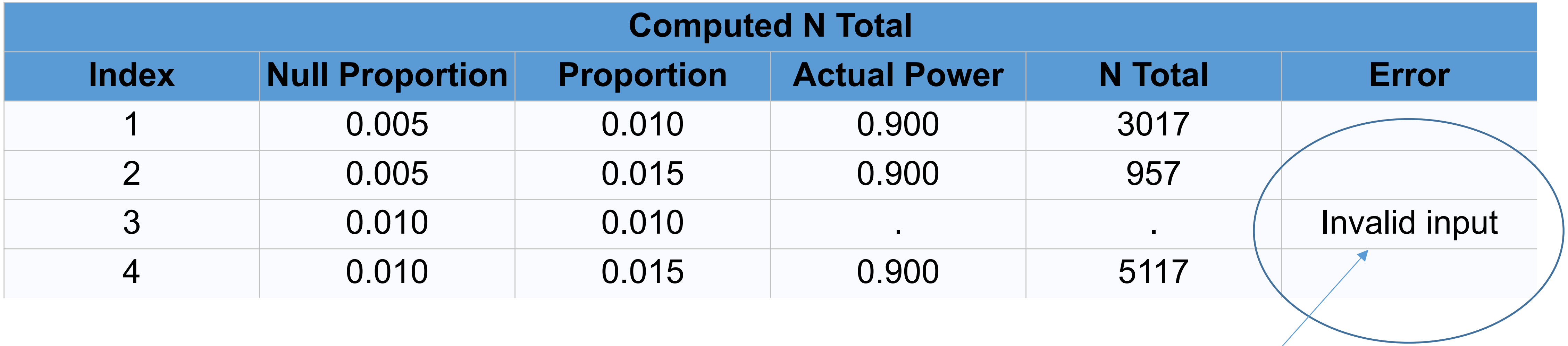
Index	Mean Diff	Actual Alpha	Actual Power	N Total
1	2	0.05	0.800	9814
2	10	0.05	0.801	396

Example 4

The color of text on a website is thought to change the behavior of customers clicking on add links. Given there are 20 million customers visiting each week and an extra 5 cents in add revenue for every click, a difference of 0.005% is considered important. What is the power needed to detect this difference if:

The conversion rate is approximately 0.5% for the ‘least popular color’

```
/*EXAMPLE 4*/
*ONE WAY TO THINK ABOUT IT IS A ONE SAMPLE TEST
*
*;
PROC POWER;
       ONESAMPLEFREQ  METHOD=NORMAL TEST= ADJZ /*SAMPLE SIZES
ARE GOING TO BE BIG WE CAN APPROXIMATE USING THE NORMAL*/
       NULLP= 0.005 TO 0.01 BY 0.005
       P = 0.01 TO 0.015 BY 0.005
       ALPHA = 0.05
       POWER = .90
       NTOTAL = .;
RUN;
END;
```



Computed N Total					
Index	**Null Proportion**	**Proportion**	**Actual Power**	**N Total**	**Error**
1	0.005	0.010	0.900	3017	
2	0.005	0.015	0.900	957	
3	0.010	0.010	.	.	Invalid input
4	0.010	0.015	0.900	5117	

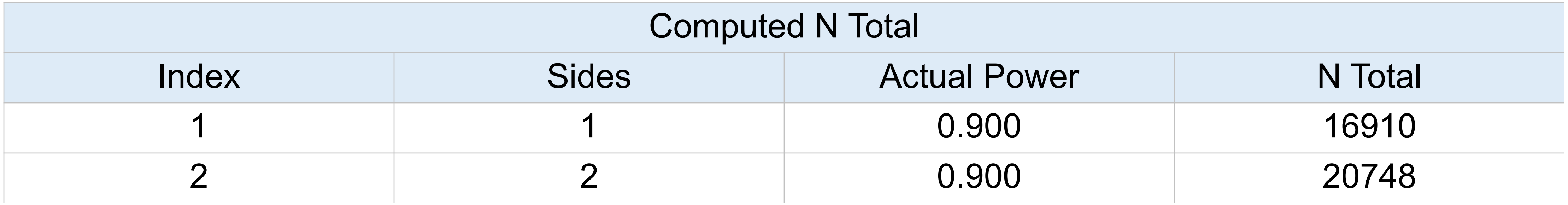
Why?

On proportions:

- The power changes with the proportion (Why?)
- Make sure you know what formula you are using when doing the comparison, this will change your power result.
- The problem above is that we are assuming one mean is known, so it doesn't really work, and will way underestimate our sample size.
- The 'twosamplefreq' will give you the correct answer.
- When you are dealing with multiple proportions you are going to want to uses something like logistic regression, but power becomes a bit more complicated as we will see in lecture 8.

THE CORRECT WAY TO DO IT:

```
/*WE HAVE THIS WONDERFUL TWO SAMPLE FREQ
  PROCEDURE CAN WE DO SOMETHING MORE CORRECT*/
PROC POWER;
       TWOSAMPLEFREQ TEST=PCHI
              refproportion=.01    /* reference
proportion is 0.01*/
              proportiondiff=.005  /* increased
conversion is 0.005*/
              sides=1 2
              alpha=.05
              power=.9
              NTOTAL=.;
RUN;
```

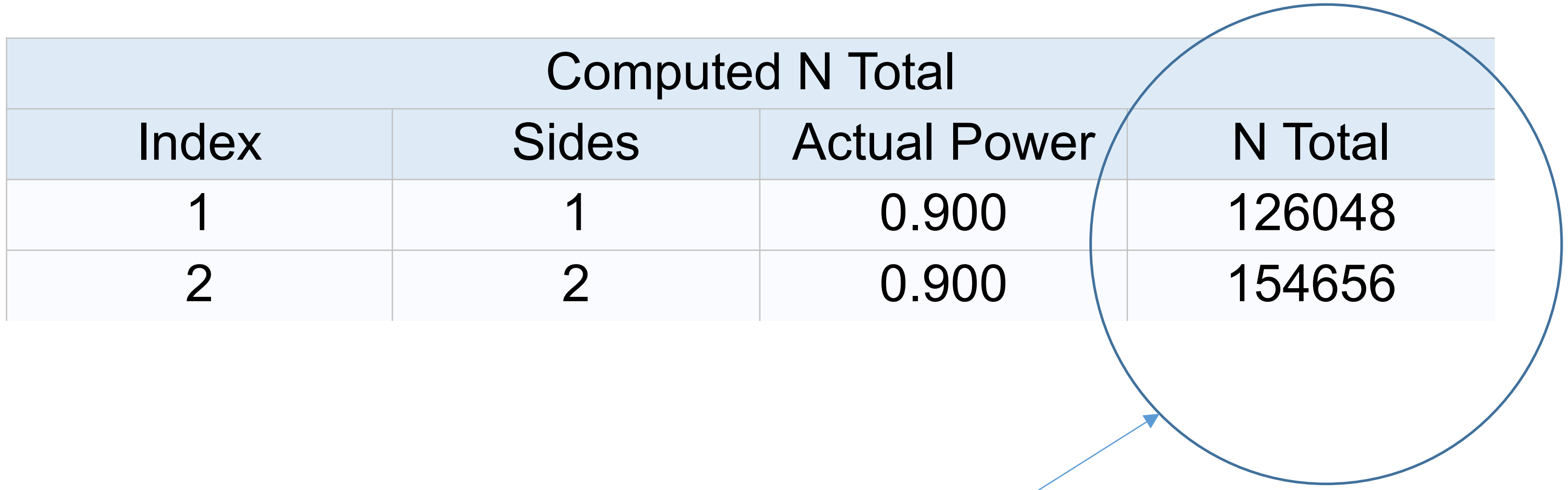


Computed N Total			
Index	Sides	Actual Power	N Total
1	1	0.900	16910
2	2	0.900	20748

The reality is we are NOT comparing one proportion to a KNOWN value so our N is much higher.

BUT A DIFFERENCE IN THE BACKGROUND PROPORTION RATE IS A BIG DEAL!

```
/*WHAT HAPPENS WHEN WE INCREASE THE REFERENCE
PROPORTION
 */
PROC POWER;
       TWOSAMPLEFREQ TEST=PCHI
             refproportion=.1    /* reference
proportion is 0.1*/
             proportiondiff=.005  /* increased
conversion is 0.005*/
             sides=1 2
             alpha=.05
             power=.9
             NTOTAL=.;
RUN;
```



Computed N Total			
Index	Sides	Actual Power	N Total
1	1	0.900	126048
2	2	0.900	154656

Now our sample sizes are almost 10x higher!

More notes on proportions:

1. Making mistakes in assumptions for normal data won’t affect your power by as much as it will a proportion.
2. Proportions variability change with the true value.

   - This variability goes down the further the proportion gets away from 0.5, which is the point of maximum variability.

3. If you can get away with not looking at the data as dichotomous,

   you are always in better shape.

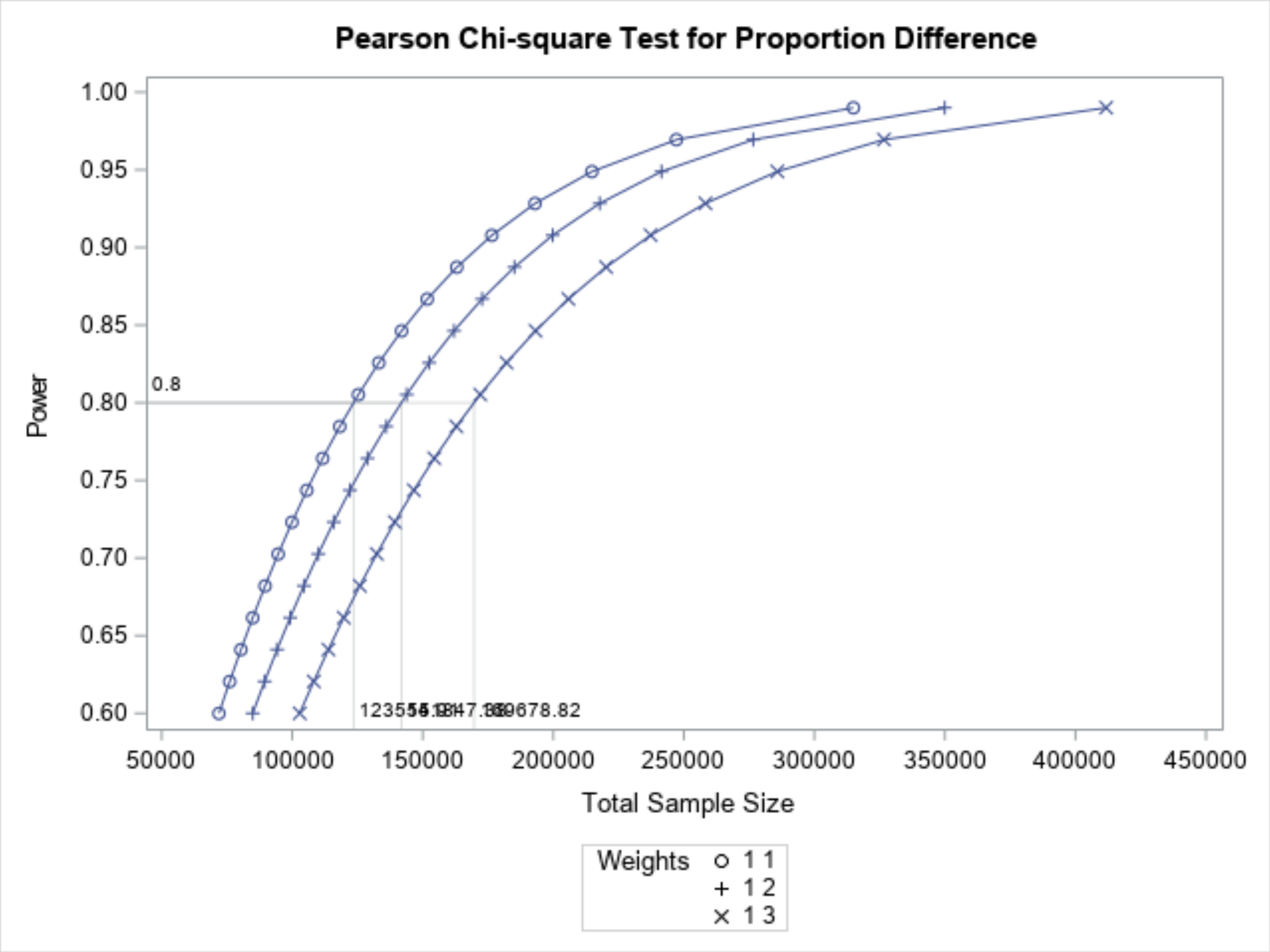
Example 5:

Whale strikes happen when boats hit a whale. When they occur it is fatal for the whale. The coast guard is considering changing its rules in a particular sound in Alaska to protect the whales, but it wants to know if its new protocol will change the risk. The probability of hitting a whale is currently about 0.001 percent with the present protocol.

Let’s say a difference of 0.0005 is important.

```
/*Example 5 HITTING A WHALE*/
ODS HTML STYLE=HTMLBLUECML;
PROC POWER;
       TWOSAMPLEFREQ TEST=PCHI
             refproportion=0.001
             proportiondiff=0.005
             sides=1
             alpha=0.05
             power=0.6 0.8 0.99
             groupweights= (1 1) (1 2) (1 3)
/*CONTROL THE SAMPLE SIZE ALLOCATION*/

             NTOTAL=.;
       PLOT Y=POWER YOPTS=(REF=0.8 CROSSREF=YES)
             VARY(SYMBOL BY GROUPWEIGHTS);
RUN;
```


Pearson Chi-square Test for Proportion Difference
Power
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.8
123556.1947.36678.82
50000
100000
150000
200000
250000
300000
350000
400000
450000
Total Sample Size
Weights
○ 1 1
+ 1 2
× 1 3

Now if there are only 10,000 boats per year we have about 10 to 20 years to see if the protocol is going to make a difference!

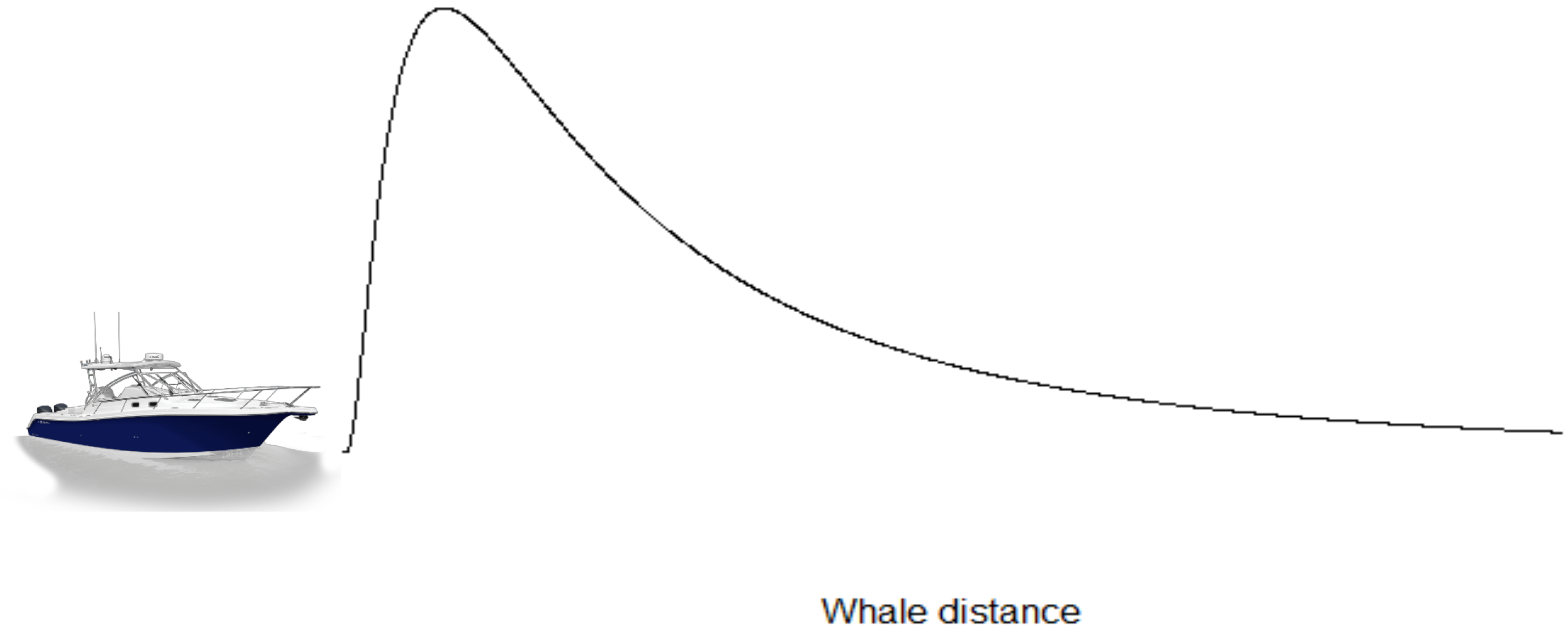
Try to convince someone of this experiment if it costs money!

Different way to think about it

Let's define a boat hit as the Whale getting below 20ft, so we can define the problem as a continuous variable. Given it is a big sound, and whales probably stay far away from boats, let's say that observations are log-normally distributed. That is the log of distances is normally distributed with standard deviation 1.

As log(20) is approximately 3. We are arguing our current log mean is 6.1 as the probability a z-score of 3.1 has the probability of 0.001.

How we can think about it visually



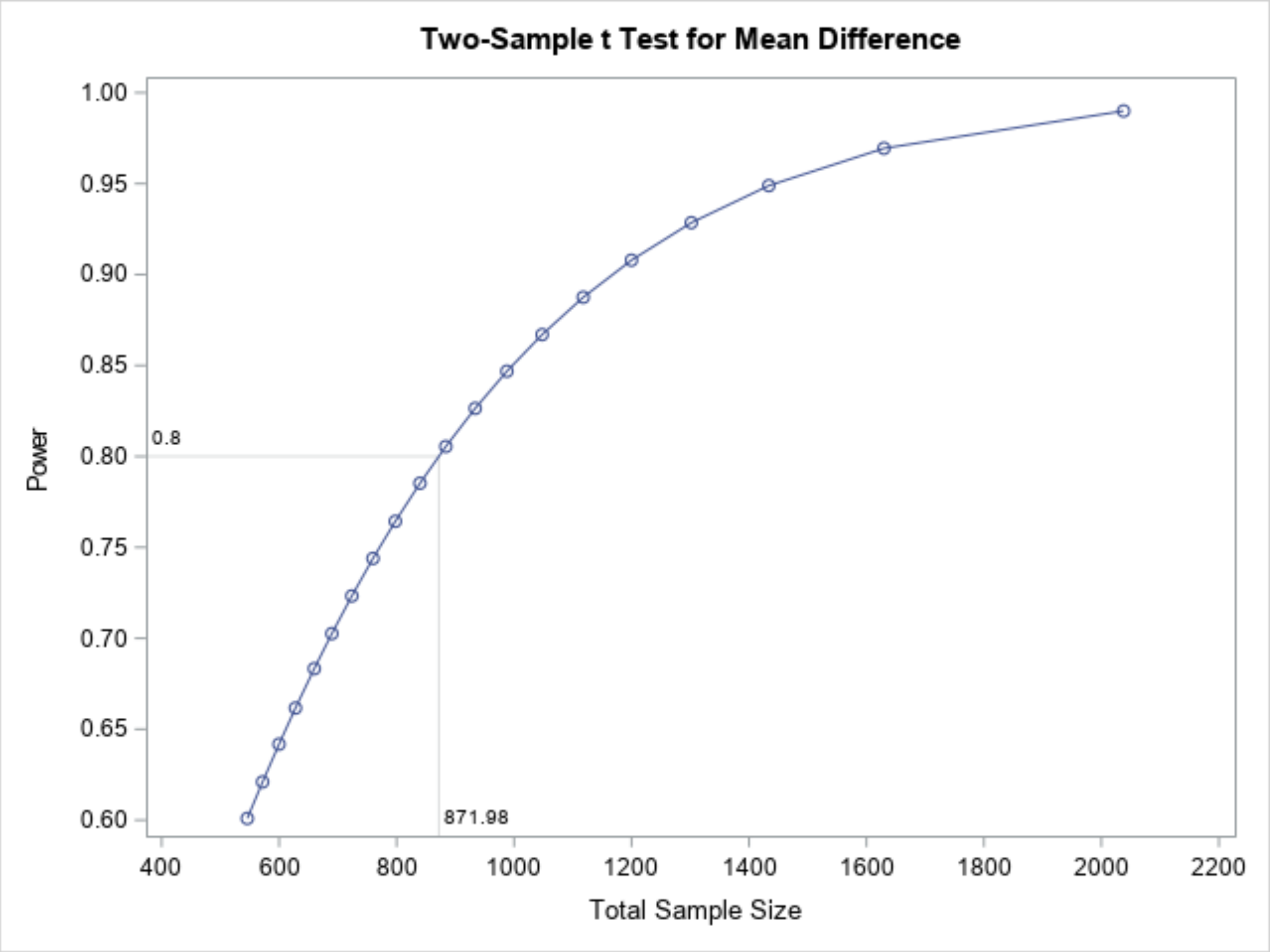
As log(20) is approximately 3. We are arguing our current log mean is 6.1 as the probability a z-score of 3.1 has the probability of 0.001.

Use pnorm(3.1) in R to see

We want to shift the log mean to have a probability of 0.0005, which implies the log mean is 6.29

```
/*Change the variable to a continuous variable
assuming the log of the distance is normally
distributed
now what do we get? */

PROC POWER;
       TWOSAMPLEMEANS
       MEANDIFF = 0.19
       STD = 1
       ALPHA = 0.05
       POWER = 0.6 .80 0.99
       NTOTAL = .;
       PLOT Y=POWER YOPTS=(REF=0.8 CROSSREF=YES);
RUN;
```


Two-Sample t Test for Mean Difference
Power
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.8
871.98
400
600
800
1000
1200
1400
1600
1800
2000
2200
Total Sample Size

Now

All you have to do is wait a month to see if the new protocol is increasing the distance and thus lowering the probability of a Whale strike.

Which one do you choose?

In an experiment, what you measure has all the difference in the world.