


How Many Observations Do I Need?

to achieve a desired precision or power

Dennis R. Helsel
Edward Gilroy

PracticalStats.com




© 2019 PracticalStats.com

1

What is Power?

		Unknown True Situation	
		H ₀ is true (no signal)	H ₀ is false (signal exists)
Decision	Fail to reject H ₀	$1 - \alpha$ Confidence Level	Type II error β (False negative)
	Reject H ₀	Type I error α Significance Level (False positive)	$1 - \beta$ Power

Power is the ability to reject the null hypothesis when it is false (can see the signal)



© 2019 PracticalStats.com

2

Two reasons to estimate power

- **Prospective study:** what power will I have with the number of observations I plan to collect? How many observations do I need to obtain 90% power?
- **Retrospective study:** I did not reject the null hypothesis. Did I have sufficient power to do so, if the groups had been different? Or do I just not have enough data to see differences?

Power and number of observations (sample size) are directly connected.



© 2019 PracticalStats.com

3

To determine sample size, you need to specify

1. alpha (α) -- tradition is to set to 5%, 0.05
2. Desired power β (tradition: I want to detect differences 90% of time, power=0.9)
3. The minimum signal (Δ) I want to be able to detect
4. The expected noise (standard deviation σ) in data

$n = \text{Function}(\Delta/\sigma, 1-\beta, \alpha)$



© 2019 PracticalStats.com

4

Minimum signal

For parametric tests, the minimum signal is Δ/σ

Δ = minimum distance between null and alternate hypotheses

σ = standard deviation (noise) of data

$$n = \text{Function} (\Delta/\sigma, 1-\beta, \alpha)$$



© 2019 PracticalStats.com

5

Approximate sample size for t-test found in most textbooks

$1-\beta$ = Power

$$n \geq \left[\frac{Z_{1-\alpha} + Z_{1-\beta}}{\left(\frac{\Delta}{\sigma} \right)} \right]^2$$

This equation is based on an assumption of normality. If non-normal, transform data first (but then it is for testing differences in medians).



© 2019 PracticalStats.com

6

Approximate Power of t-test

- Approximate because σ is unknown; s (standard deviation) is used, so t **not Z** should be used.
- But cannot use t directly because t depends on sample size n , which is unknown
- One approach is to use t, but start with an assumed n and iteratively solve for a stable n and t



© 2019 PracticalStats.com

7

Approximate Power of t-test

Second approach: increase by 3 the sample size n in each group found in the equation's approximation.

ref: Kupper, Lawrence L. & Hafner, Kerry B., 1989, The American Statistician, v. 43, no.2.



© 2019 PracticalStats.com

8

$n = \text{Function} (\Delta/\sigma, 1-\beta, \alpha)$

Δ	↑	bigger signal	n	↓
σ	↑	more noise	n	↑
$1-\beta$	↑	more sure of detection	n	↑
α	↑	significance level	n	↓



© 2019 PracticalStats.com

9

With normal theory equations, the pooled standard deviation s_{pooled} is used

Use pooled standard deviation when s is not same in both groups

$$s_{\text{pooled}} = \sqrt{\frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}}$$

This may be computed by the software and you will input s_x and s_y .

Software may or may not allow $n \neq m$ (unequal sample sizes in the two groups).



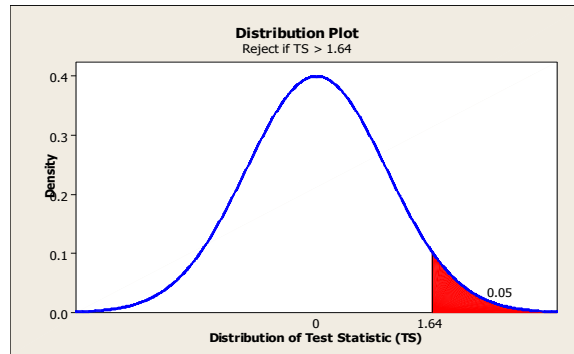
© 2019 PracticalStats.com

10

A Picture of Power

Test statistic distribution if $H_0: \mu = 0$ is true:

Red area = $\alpha = 0.05$



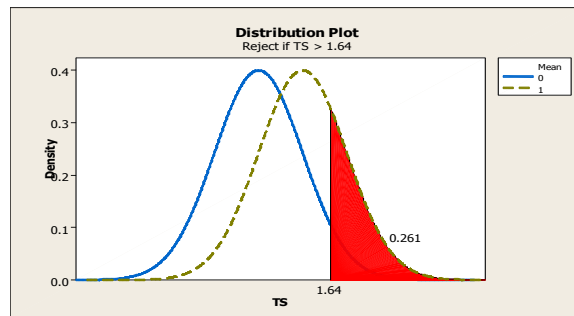
© 2019 PracticalStats.com

11

A Picture of Power

If $H_A: \mu = 1$ is true (dashed line) and $\Delta = 1$:

Power = Probability of rejection if H_A is true = red area = .261



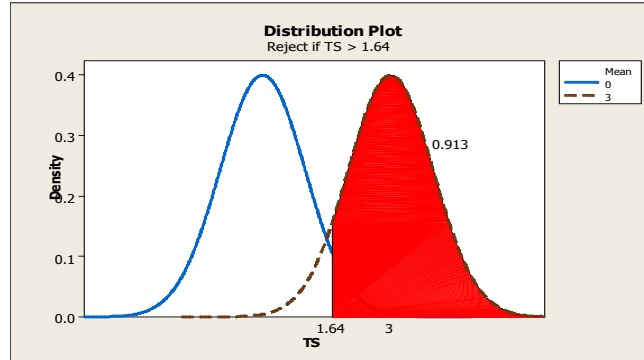
© 2019 PracticalStats.com

12

As Δ increases, power increases

Δ is larger

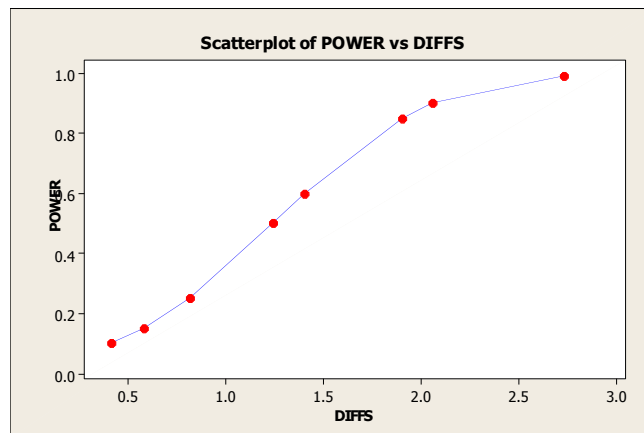
Power = Probability of rejection if $H_A: \mu = 3$ is true
= red area = .913



© 2019 PracticalStats.com

13

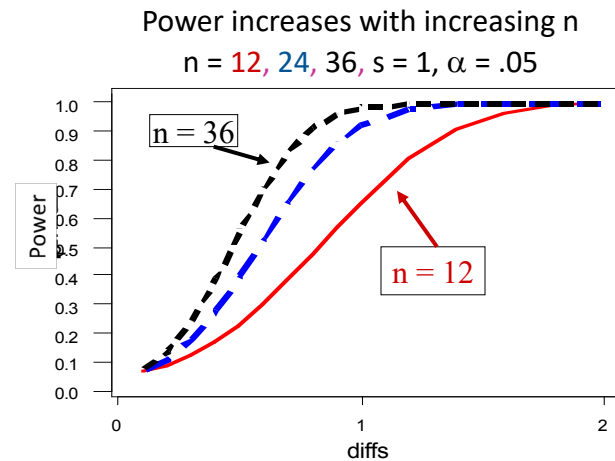
Power Curve: the red areas vs Δ/σ , the signal



© 2019 PracticalStats.com

14

Power curve, power for several sample sizes



© 2019 PracticalStats.com

15

Prospective Study

- Done before collecting data.
- Will you have enough power to detect the smallest desired difference?
- Power can be increased by
 - increasing sample size (and \$\$)
 - increasing Δ (the smallest difference of interest)
 - using a more powerful test (permutation test).



© 2019 PracticalStats.com

16

Prospective example – how many obs needed by t-test for 90% power?

```
> power.t.test (n, delta, sd_pooled, power, sig.level = 0.05,
type = c("two.sample", "one.sample", "paired"),
alternative = c("two.sided", "one.sided"))
```

{Items in purple are options: Choose one}

ONE of n, delta, sd_{pooled}, power, sig.level is not specified. It will be calculated.

Default sig.level = .05



© 2019 PracticalStats.com

17

R Example

MOLY data (compare upgradient vs downgradient concs)

Choose $\Delta = 2$ ug/L (min. difference we must be able to see)

Observed sd = 1.03 (pooled standard deviation)

$\alpha = 0.05$

Desired power = 0.9

How many observations will I need to distinguish the two groups of moly concentrations using a t-test with 90% power (assuming the data will follow a normal dist)?



© 2019 PracticalStats.com

18

Example: MOLY data

(did not enter value for n)

```
> power.t.test (delta =2, sd=1.03, power=0.9, type ="two.sample",
alt="two.sided")
```

```
Two-sample t test power calculation
  n = 6.697756 (n for each group)
delta = 2
sd = 1.03
sig.level = 0.05
power = 0.9
alternative = two.sided
```

If we collect 7 in each group, we will have 90% power to detect a difference of 2 -- IF...
 (IF data in each group follow a normal distribution, n is the same in both groups, and both groups have a std dev = 1.03). This is not what we have.



© 2019 PracticalStats.com

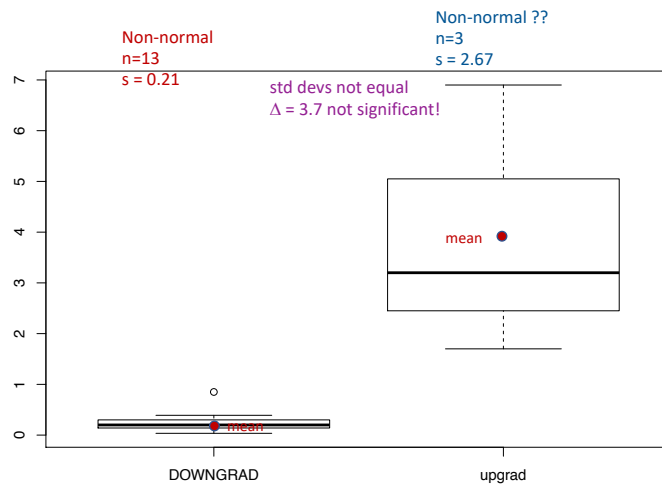
19

Reality: Meeting assumptions is important!

```
> t.test(MOLY~LOCAT)
```

Welch Two Sample t-test

```
data: MOLY by LOCAT
t = -2.3836, df = 2.0057, p-value = 0.1396
alternative hypothesis: true difference in
means is not equal to 0
95 percent confidence interval:
-10.321151  2.949254
sample estimates:
mean in group DOWNGRAD  mean in group upgrad
  0.2473846             3.9333333
```



© 2019 PracticalStats.com

20

20

Better power function: pwr package allows for unequal sample sizes

```

> 2/2.67 # d = Δ / highest sd
[1] 0.7490637
> pwr.t2n.test(n1 = 13, n2 = 3, d = 0.75,
sig.level = 0.05, alt = "two.sided")

t test power calculation
n1 = 13
n2 = 3
d = 0.75
sig.level = 0.05
power = 0.1937942
alternative = two.sided

Observed power = 19%.
Still is assuming normality
and equal sd (though we did take
the worst case). We will need to
collect more data.
    
```

© 2019 PracticalStats.com 21

21

Better power function: pwr package What n is needed for 90% power?

```

> pwr.t2n.test(n1 = 30, d = 0.75, sig.level =
0.05, power = 0.9, alt = "two.sided")

t test power calculation
n1 = 30
n2 = 52.83201
d = 0.75
sig.level = 0.05
power = 0.9
alternative = two.sided

To get 90% power will require
something like 83 total observations.
Still is assuming normality and equal sd
    
```

© 2019 PracticalStats.com 22

22

Retrospective Study

- After collecting data and H_0 is not rejected.
- What was the probability of detecting the difference (effect) you wished to detect? -- assuming the data are normally distributed?
- Did you have enough observations to see the desired difference?
- Often collect a small amount of data, compute sd and see how much additional data must be collected



© 2019 PracticalStats.com

23

Retrospective Study. What is the power I achieved?

- Specify desired α
- Specify n , s , Δ from the data you've collected
- Estimate the power (probability of detecting a difference) that you achieved with that sample size.



© 2019 PracticalStats.com

24

Assuming normality and equal sd, but correcting for unequal n:

Observed $\Delta = 3.7$.

```
> 3.7/2.67 # d = observed  $\Delta$  / highest sd
```

```
[1] 1.386
```

```
> pwr.t2n.test(n1 = 13, n2 = 3, d = 1.386,
  sig.level = 0.05, alt = "two.sided")
```

```
t test power calculation
```

```
n1 = 13
```

```
n2 = 3
```

```
d = 1.386
```

```
sig.level = 0.05
```

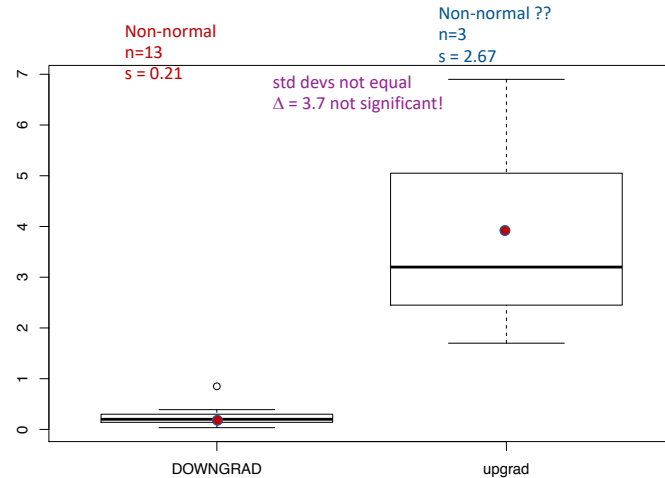
```
power = 0.5218151
```

```
alternative = two.sided
```

says we have 52% power.
still is assuming normality
and equal sd, so actual power
will be lower than this.



© 2019 PracticalStats.com



25

Retrospective Study

If you failed to reject H_0 and the power is low (< 0.70):

1. Modify your design
2. Collect more data
3. Stop the study – not worth the money
4. Use a different test – permutation?
nonparametric?
take logs -- test medians?



© 2019 PracticalStats.com

26

Transformations for non-normality. Power?

- Test difference in ratio of geometric means by taking logs.
- For retrospective, compute the observed difference in mean logarithms and divide by highest observed std dev to use as Δ
- What power did we achieve for a test of difference in geometric means (approx. medians) ?



© 2019 PracticalStats.com

27

Power for testing difference in medians, assuming normality and equal sd of logarithms:

Observed mean1 - mean2 = 2.9

```
> 2.9/0.87 # d = observed Δ / highest sd
```

```
[1] 3.333333
```

```
> pwr.t2n.test(n1 = 13, n2 = 3, d = 3.33,
sig.level = 0.05, alt = "two.sided")
```

```
t test power calculation
```

```
n1 = 13
```

```
n2 = 3
```

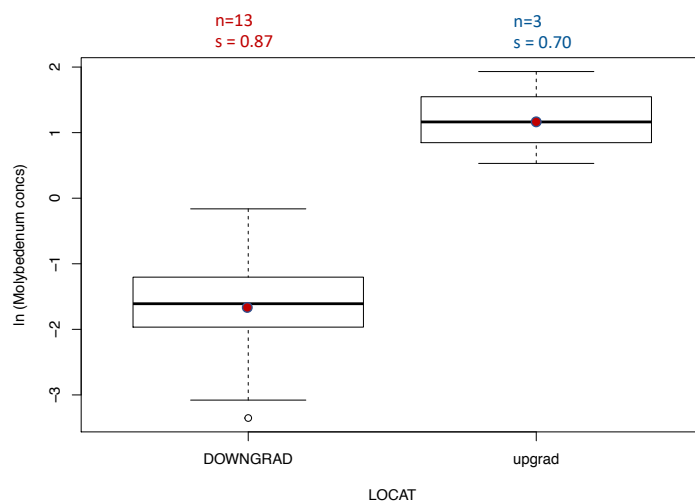
```
d = 3.33
```

```
sig.level = 0.05
```

```
power = 0.9978635
```

```
alternative = two.sided
```

says we have 99% power.
assuming normality and equal sd
of logarithms. Actual power will
be slightly lower.



© 2019 PracticalStats.com

28

28

Retrospective Study

If you failed to reject H_0 and the power is high (say 0.9):

1. You have shown there is no meaningful difference (effect). You are done.



© 2019 PracticalStats.com

29

Parametric tests: power and sample size. R commands

power.t.test **standard two-group t-test**

power.t.test "one.sample" **t-confidence intervals**

power.t.test "paired": **paired t-test**

power.anova.test **one-factor ANOVA**

SamplingStrata package **multi-level nested ANOVA**

pwr package **t-tests, ANOVA with unequal sample size**

**** recommended**

All are based on a normality assumption.

All require estimates of σ , Δ , α , and power or n



© 2019 PracticalStats.com

30

Are there sample size and power analyses for nonparametric tests?

YES

REFERENCE:

"Sample size determination for some common nonparametric tests" by Gottfried E. Noether (1987), Journ. Of American Statistical Assoc., v. 82, no. 398, p. 645-647.



© 2019 PracticalStats.com

31

Minimum signal Δ for nonparametric tests

For nonparametric tests, the minimum signal Δ is Pplus

Pplus = probability that for any random Y and X, $X > Y$.

Pplus for null hypothesis $H_0 = 0.5$

$n = \text{Function}(\text{Pplus}, 1-\beta, \alpha)$



© 2019 PracticalStats.com

32

$P_{plus} > 0.5$ is the evidence for the alternative hypothesis

Null Hypothesis: $\text{Prob}[x > y] = 1/2$
 $P_{plus} = 1/2$

Alt. Hypothesis: $\text{Prob}[x > y] \text{ NOT} = 1/2$
 $P_{plus} = 0.8$ (for example)



© 2019 PracticalStats.com

33

Nonparametric tests: power and sample size. R commands

`power.prop.test` test for two-group difference in proportions
`wmwpow` package rank-sum test

Practical Stats R scripts for power of nonparametric tests:

`pownp` rank-sum, signed-rank and sign tests -- provided with the
 online Applied Environmental Statistics course.

`power.WMW` rank-sum test -- provided with Statistical
 Methods in Water Resources 2nd Edition.

All require estimates of Δ ($\text{prob } x > y$), α , and power or n



© 2019 PracticalStats.com

34

If Pplus doesn't seem intuitive:

$$\begin{aligned} P_{\text{plus}} &= \text{Prob}(x > y) \\ &= \text{Prob}(\ln x > \ln y) = \text{Prob}(\ln x - \ln y > 0) \end{aligned}$$

IF the logarithms can be assumed to follow a normal dist,
Pplus can be retransformed back into the ratio of geometric means.

$$= \text{Prob}(\text{geomean}(x) / \text{geomean}(y) > 1)$$

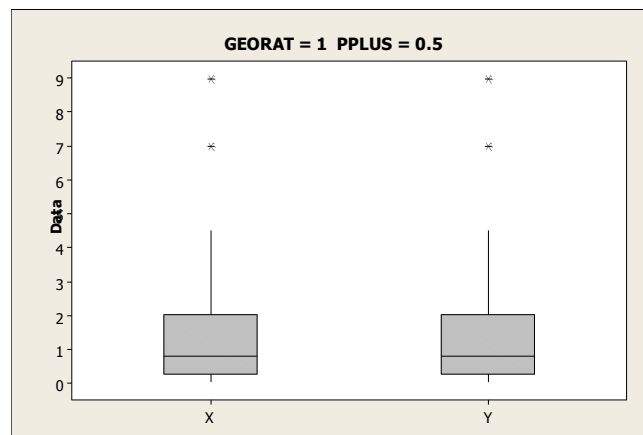


© 2019 PracticalStats.com

35

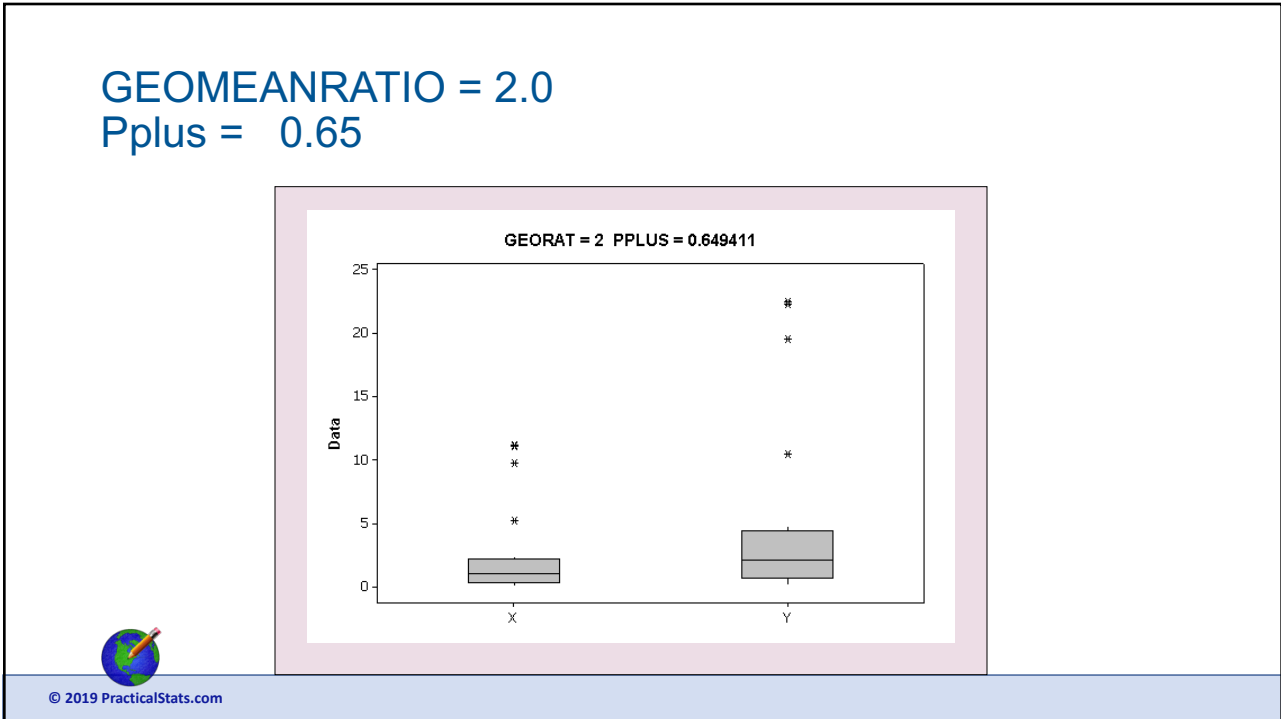
GEOMEANRATIO = 1.0

Pplus = 0.5
(the null hypothesis)

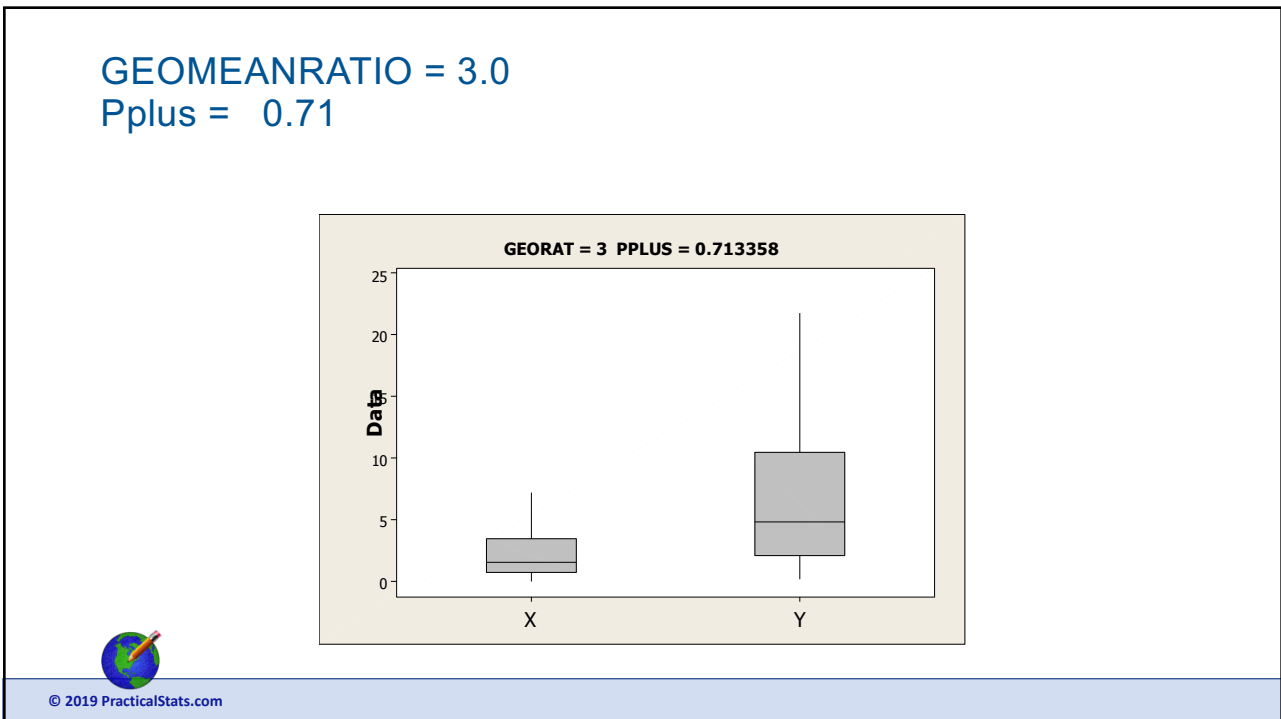


© 2019 PracticalStats.com

36

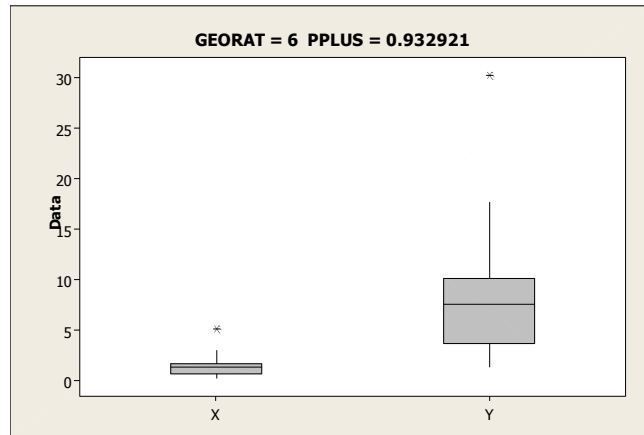


37



38

GEOMEANRATIO = 6.0
Pplus = .93

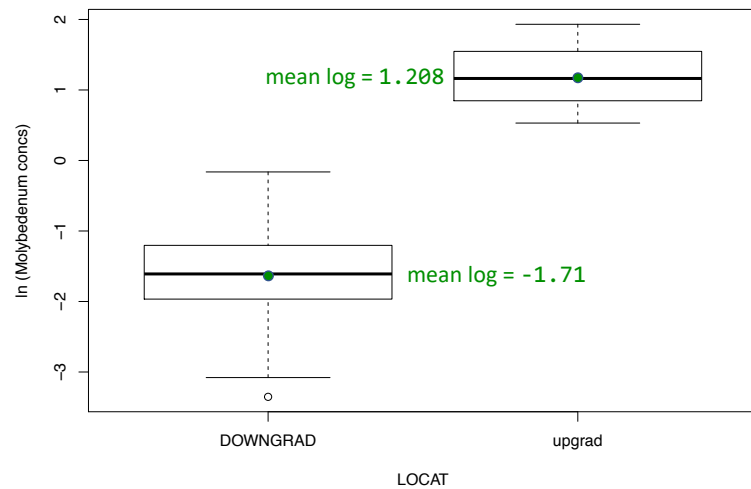


© 2019 PracticalStats.com

39

For the Moly data:

Ratio of Geomeans = 18.51
Pplus = 0.995
Std Dev (logs) Upgradient = 0.868
Std Dev (logs) Downgradient = 0.701



© 2019 PracticalStats.com

40

power.WMW script

```

> geomean1 = exp(-1.709755)
> geomean2 = exp(1.208433)
> georatio = geomean2/geomean1
> georatio
[1] 18.50772


> power.WMW(MOLY, LOCAT, gmratio = 18.5, conf = 95)
SampleSize is the required number of obs in both groups together.
Nratio is the proportion of SampleSize for 1st variable entered. 13/16 = 0.8125
SampleSize Nratio GMratio PPlus ObsrvPower
1          16 0.8125 18.5 0.996      84.97
    
```

Sample sizes are rounded up to smallest integer not less than the computed sample size

SampleSize	Power
7	55.1
7	55.1
9	64.2
10	68.2
11	71.8
12	75.0
14	80.6
17	86.8
20	91.2
25	95.6 (n = 20, m = 5)

We achieved ~85% power with 3 and 13 observations.

Pplus (Prob X>Y) was 99.6. Can't do much better.



© 2019 PracticalStats.com 41

41

Power to detect differences in MOLY groups, retrospective, accounting for unequal n in groups. Depends on measure of difference.

NUMBER OF OBSERVATIONS TOTAL	t-test on means (IF each group follows a normal distribution)	t-test on logs of MOLY (tests diffs in geometric means)	Rank-sum test of diffs in medians (0.81*N in one group)
16	52%	99%	85%

↑


Too high because data NOT normal and sd are not equal!

↑

Realistic if geomeans wanted because logs of data close to normal ! Assumes groups follow lognormal distribution

↑

Realistic. No normality or equal sd required.



© 2019 PracticalStats.com

42

Other Newer Approaches

Permutation Methods

- Use permutations of existing data after subtracting off difference in means/medians/geometric means
- Add in the desired signal
- Run 10K times or so, and record percent of times the signal is detected (**Power**) using a permutation test (not t-test)
- Change the sample size and do it all again (**get power curve**)
- Pros: No assumptions of distribution / equal variance / equal n are needed
- Con: Software is 'build your own'



Commercial package for sample size computations -- PASS

Confidence intervals	signed-rank test	rank-sum test
Tolerance intervals	Correlation	Kruskal-Wallis test
Tests of proportions	Contingency tables	
t-tests	ANOVA	Repeated measures
Mixed models	Factorial ANOVA	Tests of variances

* Nonparametric methods

* Parametric methods

Note: I have no connection with PASS software and have not used it.



Free online Sample Size calculator for parametric methods

<https://samplesizeshop.org/>
Univ. of Colorado


allows for unequal sample sizes. All methods assume a normal distribution.

The use of this site is free and always will be.

We provide researchers such as behavioral and social scientists with tools and education related to the following aspects of study design:

- Calculating power and sample size for any univariate or multivariate test for the general linear multivariate model, assuming fixed predictors.
- Producing confidence intervals on power estimates for designs with fixed predictors.
- Producing power calculations for designs with a single Gaussian covariate.
- Supporting designs with unequal group sizes, and complicated covariance structures.
- Creating basic power curves.

To calculate power or sample size, click on the Home tab above and then on the orange 'Calculate Sample Size Now' button. For educational materials related to study design, please peruse the papers found under the 'Resources' tab.




© 2019 PracticalStats.com 45

45

Next Month's Webinar

Tuesday March 17th 11 am Mountain time

- **Incorporating Greater Than and Less Than Values In Data Analysis**
Tuesday Mar 17, 2020 1 pm Eastern, 10 am Pacific
<https://attendee.gotowebinar.com/register/884495792221105163>
- Sign up for our newsletter/announcement list to get the registration link emailed to you. <http://practicalstats.com/news/>
- Or check our webinars page periodically at <http://practicalstats.com/training/webinar.html> to register for it.



© 2019 PracticalStats.com 46

46

Today's webinar will be available soon for streaming

on our new Videos page

<http://practicalstats.com/videos/>

Let colleagues who missed it know about it.



47

Thank you for attending

- Much of the material is based on Chapter 13 of the book [Statistical Methods in Water Resources, 2nd Edition](#) by Helsel, Hirsch, Archfield, Ryberg and Gilroy (2020).
- All opinions are my own and do not represent those of anyone else.
- Questions?

Get in touch!

Dennis Helsel ask@practicalstats.com

Courses & free recordings at: <http://practicalstats.teachable.com>



48