Sample Size Estimation and Power Computation on Paired or Skewed Continuous Data

March 31, 2006
GCRC Weekly Workshop
Ayumi Shintani, Ph.D, M.P.H.

Departmant of Biostiatistics
Vanderbilt University Medical Center

Terminology

Ho: $\mu=0$ versus H1: $\mu \neq 0$. Where μ = mean % change in level of pain over 1 month period.

Type 1 error (α): the probability of deciding the drug is effective, given the true state of nature is that the drug has no effect on pain relief.

Type 2 error (β): the probability of deciding the drug has no effect on pain relief based, given the true state of nature is that the drug is an effective pain reliever.

Power of the test ($1-\beta$): the probability of deciding that the drug is effective pain reliever, when the true state of nature is that it is effective.

General Concepts of Sample Size and Power

The calculation of power is used to plan a study, usually before any data have been obtained, except possibly from a small preliminary study called a pilot study.

Recall:

The general aim in hypothesis testing is to use statistical test that make both Type I and Type II error small. Traditionally, a Type I Error is fixed at 0.05, and we try to get a sample large enough to ensure that 1-ß is at a reasonable level (> 0.80).

Factors Affecting the Sample Size:

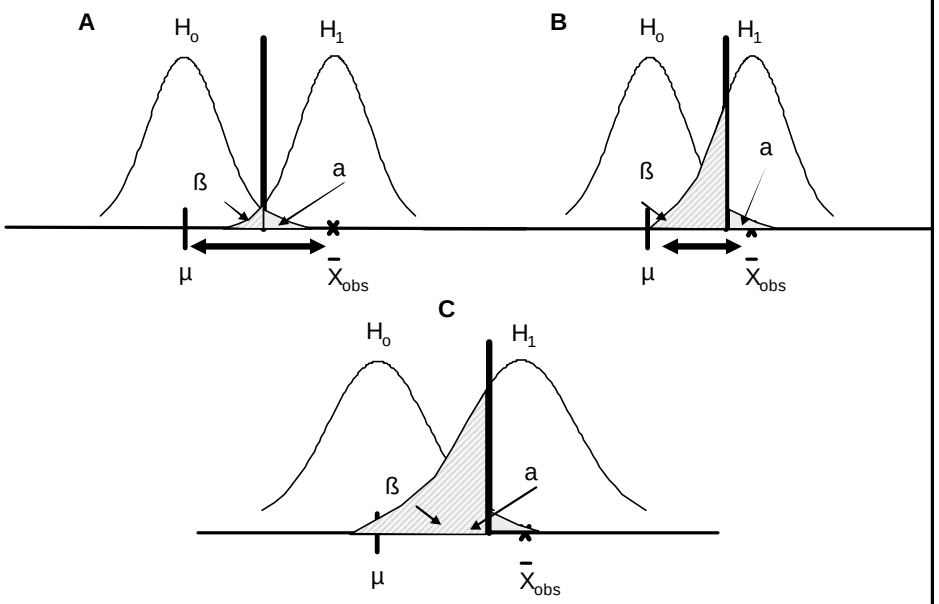
(1) The sample size decreases as the difference between the null and alternative means increases (effect size). A vs B on the following page figures

(2). The sample size increases as standard deviation (SD) increase, i.e., more variability of data, sample size increases. A vs C on the following page figures

(3). The sample size increases as the significance level is made smaller (a decreases). Usually fixed at two-sided 0.01 or 0.05

(4) The sample size increases as the required power increases (1-ß increases). Usually targetted at 80 or 90%.

Effect size, and SD are usually obtained through pilot studies, or published data.



What do you see from these 3 figures?

Common approaches to sample size and power analysis

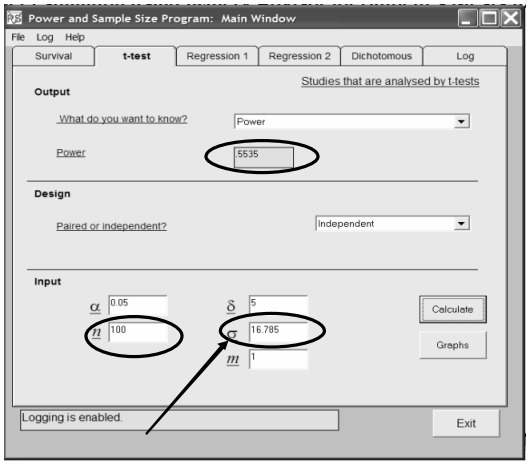
Given / Targeted		What to estimate
Power, Effect Size, SD	→	How many samples required
Sample size, Effect Size, SD	→	How big is the power
Sample size, SD	→	What is the minimum detectable difference to achieve at least 80% power
Sample size	→	What is the minimum detectable difference (in term of SD) to achieve at least 80% power

Review: Estimation of Power for Student's t-test (two sample t-test) (Rosner page 333)

Suppose 100 OC users and 100 non-OC users are available for study and a true difference in mean SBP of 5 mm Hg (132 vs 127) is anticipated, with OC users having the higher mean SBP. How much power would such a study have assuming that the variance estimates in the pilot study are correct.

Pilot study data:
OC users: mean SBP= 132.44 mm Hg with SD=15.34 mm Hg.
Non-OC users: mean SBP=127.44 mm Hg with SD=18.23 mm Hg.

Power estimation using PS for Student's t-tests (3)



We are assuming two samples are from normal distribution with the same standard deviation, thus we need to use a pulled sample standard deviation for two equal samples, since sample size is equal between 2 groups, this is an average of two standard deviation. 16.785=(15.34 + 18.23)/2

Computing Pulled SD for Student t-test

We now want to estimate s by sample standard deviation, s.

The pooled estimate of the variance from two independent samples is given by

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

When s_1^2, s_2^2 are sample variances for sample 1 and sample 2, the average of s_1^2, s_2^2 could simple be used to estimated s^2.

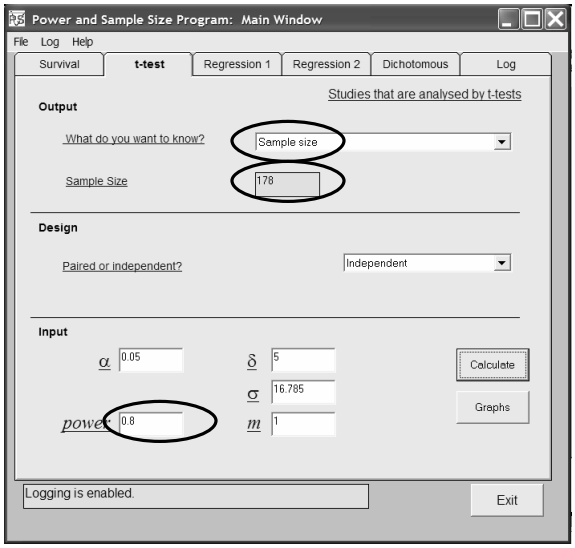
However, this average will weight the sample variance equally even if the sample sizes are very different. Similar to mean, more the sample size, estimation of population standard deviation becomes more accurate.

3.3.2. Estimation of sample size for Student's t-test (independent sample t-test) (Rosner page 333)

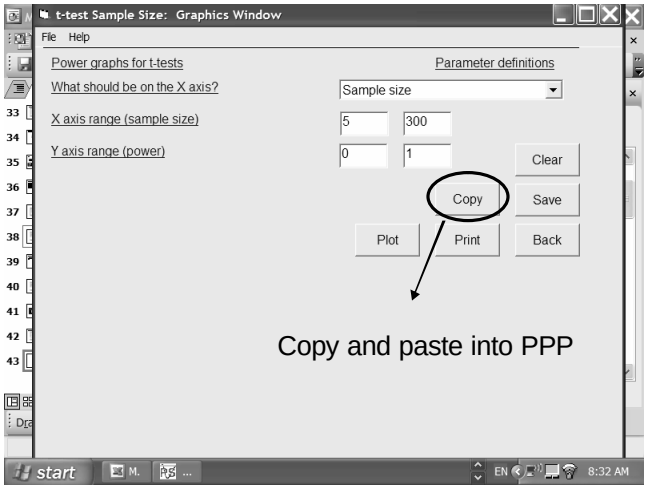
The previous power estimation shows that the anticipated power for the analysis with 100 OC users and 100 non-users is 55%. We usually needs at least 80%, (or 90% is even better). We wish to test the hypothesis Ho: $\mu_1 = \mu_2$ vs H1: μ_1 ? μ_2.

How many samples are required to achieve 80% power?

Sample size estimation using PS software for Student's t-tests (1)

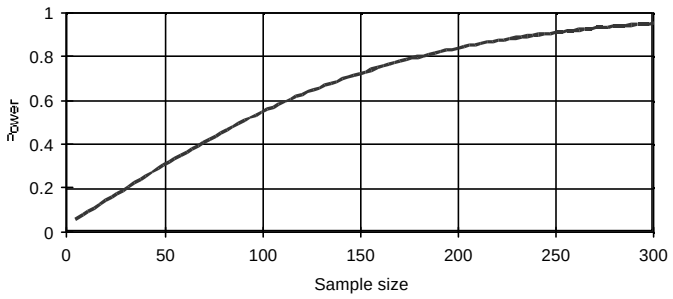


Sample size estimation using PS software for Student's t-tests (2)



Sample size estimation using PS software for Student's t-tests (3):

Power – sample size graph from PS software.



It requires about 138 patients to achieve 80% power.

Power and Sample Size Estimation for paired t-test

Review of Paired t Test:

These 2 observations are related, since they came from the same person.

SBP levels (mm Hg) in 10 women while not using OC (baseline) and while using (follow-up) OC.

i	*SBP level While not using OC*	*SBP level While using OC*	*Difference*
1	115	128	13
2	112	115	3
3	107	106	-1
4	119	128	9
5	115	122	7
6	138	145	7
7	126	132	6
8	105	109	4
9	104	102	-2
10	115	117	2

To compare the difference between the BP of women between OC users and non-users, we use the within patient difference between baseline and follow-up BP.

- If the mean of the difference = 0, we can conclude there is no difference.
- If the mean of the difference (baseline – follow-up)> 0, we can conclude that non-users have a higher BP than users.
- If the mean of the difference (baseline – follow-up) < 0, we can conclude that non-users have a lower BP than users.

Ho: $\mu_d \neq 0$. Where μ_d = the difference between baseline and follow-up values

Thus sample size and power computation requires SD of within patient difference.

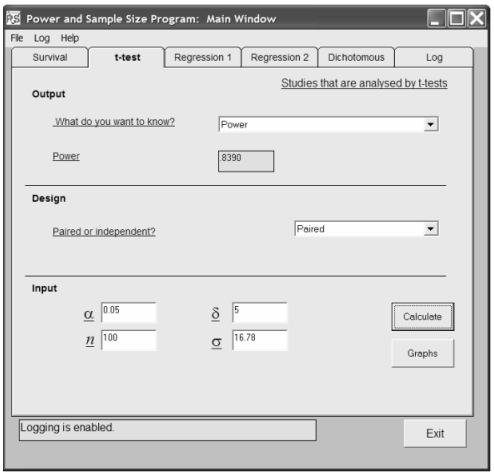
Power and Sample Size Estimation for paired t-test (for longitudinal / follow-up study).

Suppose we are planning a longitudinal (follow-up) study including 100 women to compare the mean change in SBP between baseline (all women were non-OC users) and 1 year follow-up when all women become OC users).

Previous data [This is the same as what we used for Student t-test example]: OC users (follow-up): mean SBP= 132.44 mm Hg with SD=15.34 mm Hg. Non-OC users (baseline): mean SBP=127.44 mm Hg with SD=18.23 mm Hg. Pulled SD = 16.785, effect size=5

We need to use standard deviation for this difference. But pilot data do not provide it, so we will assume the SD for within patient difference is the same as the pulled SD from baseline and follow-up.

Commonly done: Power estimate with the same SD for within patient difference



Which provides power of 84% ⟶ Is there any way to improve this?

To improve the previous power analysis, we in fact want more accurate estimate of SD for the within patient difference.

Variance of a difference in SBP = Var(x1-x2)=var(x1) - 2cov(x1, x2)+var(x2).
Where x1 is SBP from baseline, x2 is SBP from follow-up.

$$s^2_d = s^2_1 + s^2_2 - 2r s_1 s_2$$

Where
s^2_1=variance of baseline values within a treatment group
s^2_2=variance of follow-up values within a treatment group
r =correlation between baseline and follow-up values over time within a treatment group. In this example we assume r =0.7

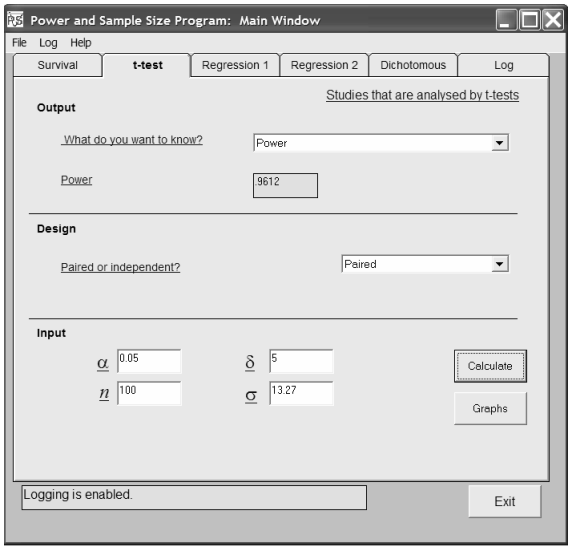
Thus, variance of the difference
= $15.34^2 + 18.23^2 - 2x0.7x15.34 \times 18.23 = 176$
Estimated SD for the difference = $\sqrt{176} = 13.27$

This is what you are going to use in PS

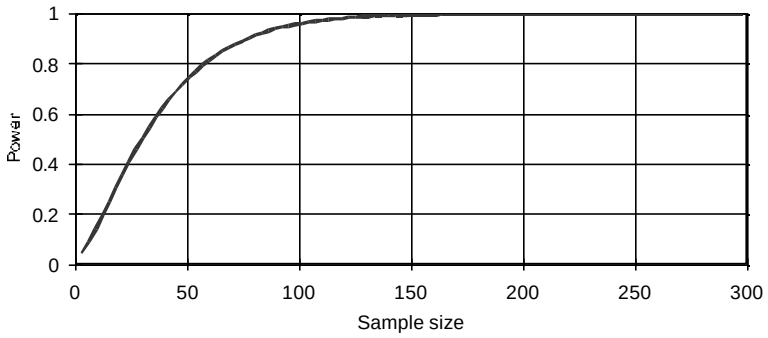
If you have a raw data which you can directly obtain SD for the difference you don't need to go through the above computation.

Power analysis we conducted on the page 17 in fact assumed that correlation between baseline and follow-up measurements is 50%, which made SD for within patient difference same as the pulled SD from baseline and follow-up.

Power estimate with baseline-followup correlation = 0.7



Power and estimation for paired t-test using nQuery Advisor (2).



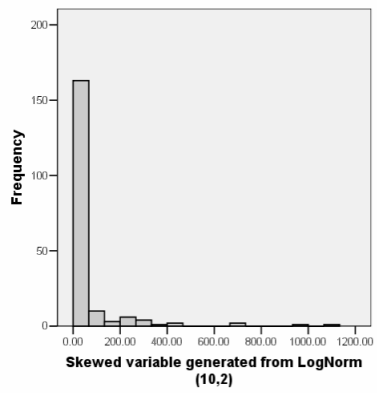
As you see, the two sample t-test example shows that the power of n=100 (200 total) is 55%, and the power of n=100 (100 total) for paired t-test is 96%.

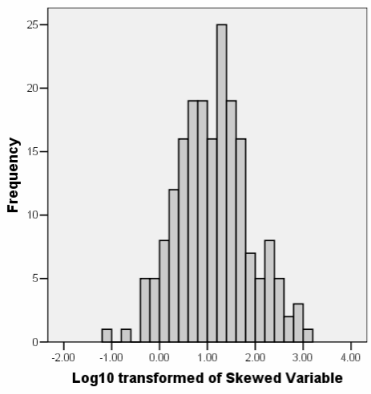
Paired analysis is very powerful, the power increases (and required sample size decreases) as correlation between baseline and follow-up variable increases.

Whenever you design an experiment, you may want to consider measuring baseline values for an outcome variable, so that you can perform more efficient analysis.

This means that when you conducted a study where measurements are taken twice from a patient, which requires paired t-test, instead, if you perform two sample t-test, you will be penalized for choosing a wrong analysis by losing a tremendous power for the analysis, which makes you harder to detect a statistical difference.

Power and sample size estimation for skewed variable.

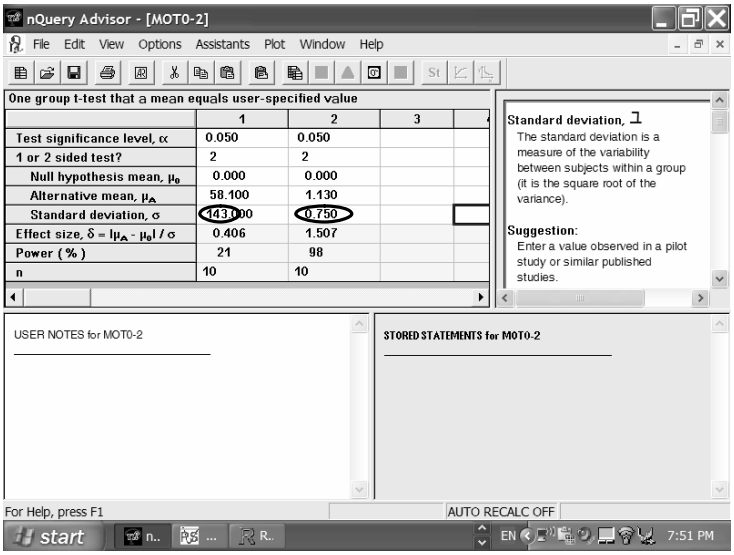




Mean=58.1
SD=143.17
N=193

Mean=1.13
SD=0.75
N=193

Power and sample size estimation for skewed variable (N=10)



As you see in this example properly performing an analysis (via variable transformation) by carefully examining distribution of variable, you will gain not only through making a right inference, but you will gain power to detect statistically significant difference.

This means that when you estimate power or sample size for obviously skewed variable, you may consider transformation of skewed variable for the estimation. You have two ways to do this:

(1) Transform row data, then use mean and SD of the transformed data for power/sample size computation.

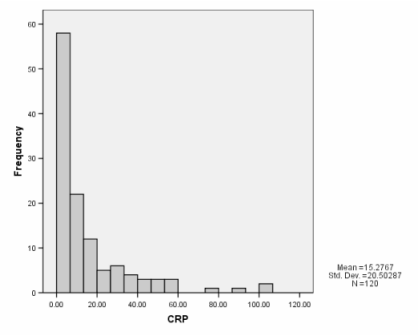
(2) Transform published mean and SD for power/sample size computation. Use this option, only when you don't have access to a row data.

Many paper present median and inter-quartile rage (IQR) for skewed data, you can estimate SD from IQR by using the following formula,
$SD=\{(IQR)/2\}/z_{0.25}=\{(IQR)/2\}/0.675$
However this is not very wise, since people often show IQR for skewed data. Estimated SD matches with IQR only for normally distributed data.

Example:

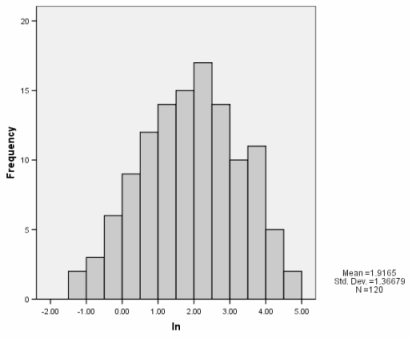
In our preliminary data from 117 patients, mean (SD) CRP was 15.3 (20.5). With anticipated 20% reduction (15.3 for control and 12.2 for intervention group), with 10 patients in each group (a total of 20), power would become 2% at two-sided significant level of 0.05.

Pilot data (not transformed)



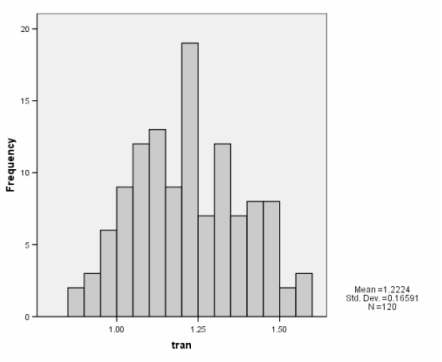
In our preliminary data from 117 patients, CRP data are highly skewed thus we attempted transformation to achieve normality. Mean (SD) of transformed CRP was 1.91 (1.37). With anticipated 20% reduction (1.53 for intervention), with 10 patients in each group (a total of 20), power would become 4% at two-sided significant level of 0.05.

Pilot data ($\log_e$ transformed)



In our preliminary data from 117 patients, CRP data are highly skewed thus we attempted transformation to achieve normality. Mean (SD) of transformed CRP was 1.22 (0.166). With anticipated 20% reduction (0.98 for intervention), with 10 patients in each group (a total of 20), power would become 76% at two-sided significant level of 0.05.

Pilot data (10^{th} root)



Final statement included in this grant application

There is no data available to estimate SD for the within patient change score thus we used CRP baseline data of 112 patients from one of our preliminary data by assuming SD for the change would be similar to SD from the baseline CRP. Mathematically, they become equivalent when within patient correlation between baseline to 1 month measurement is 50%. We anticipate within-patient correlation would be higher than 50% thus the proposed study would provide even higher power than our estimated value because higher within-patient correlation provides smaller SD. In our preliminary data from 117 patients, CRP data are highly skewed thus we attempted transformation to achieve normality. Mean (SD) of transformed CRP was 1.22 (0.166). With 10 patients in each group (a total of 20), the minimum detectable difference between control and intervention is 18% (1.22 for control group, and 1.00 for the intervention group).