# CLINICAL EPIDEMIOLOGY AND POPULATION HEALTH

## Key Points – Power and Sample Size Estimation

I. **Power** is the chance that you will detect a difference in an outcome (i.e. reject the null) when the alternative hypothesis is true in your study data (i.e. a difference in the outcome truly exists). As you will read below, power depends on the size of the difference you want to detect, the type 1 error you set, and the characteristics of your sample.

II. **Type I error** is the probability of concluding, based on a sample, that a difference exists between the groups (that they are drawn from two different source populations), when they are, in fact, the same (drawn from a single underlying population). The probability of a type I error is $\alpha$ (alpha). Alpha is the threshold probability (set prior to analysis, often by convention at .05), below which we consider the observed between-group difference to be "unlikely" due to chance. A test done with alpha = 0.05 will reject the null hypothesis of no difference 5% of the time when the truth is that there is no difference. A test achieves "statistical significance" when the p-value is less than alpha.

**III.** **Type II error** is failure to detect (come to the conclusion) that a difference exists between groups, when a difference truly exists (they are drawn from different underlying populations). β (beta) is the probability of Type II error. The complement of β is power (1–β). If some specific $H_A$ (*alternative hypothesis*) is true in the population, the *power* is the chance that you will detect a difference (reject the null) when you collect data on your *sample*. Note, however, that then rejecting the null hypothesis based on your sample does not support any specific $H_A$.

**IV. Power calculations are performed before the study is started to:**

A. Determine the sample size required to have adequate power to detect a pre-specified difference – this should be a clinically meaningful difference
B. Or, (if the sample size is fixed) to determine the size of the difference you will be able to detect with a given power.
C. Or, if the sample size and effect size are fixed, the Type II error you will have.

**V. Calculating Power is important because:**

A. If too few subjects are enrolled, the power to detect real and important differences may be small, i.e., there might be a large probability that the null hypothesis is not rejected, even when it should be.
   1. Money and time are wasted and subjects are exposed to an experiment that is inconclusive.
   2. Value of small trials could be redeemed by subsequent incorporation into meta-analysis.
B. If too many subjects are enrolled:
   1. Money and time are wasted, and some subjects are exposed to the experiment unnecessarily.
   2. The difference is measured with more-than-necessary precision.
   3. We may detect differences that are statistically significant but so small in absolute magnitude that they are not really of interest (i.e. not clinically meaningful)

If a study has a null result, it should be interpreted in the context of its power for finding a difference of a particular magnitude so that readers know what the detectable difference was at the outset of the study.

## VI. Power depends on the following components:

**Sample size:** Larger n = more power (n impacts the standard error of the groups you are trying to compare – the tighter the distribution, the easier it is to detect a difference)

**Spread of the data:** Smaller standard deviation = more power (SD impacts the standard error of the groups you are trying to compare – the tighter the distribution, the easier it is to detect a difference)

**Difference you wish to detect:** Larger magnitude of detectable difference = more power

**Willingness to commit a Type 1 error:** Larger alpha = more power. Note that there is a trade-off between committing a Type 1 error (false positive) and a Type 2 error (false negative). While researchers typically set the Type 1 error to alpha=0.05 (i.e., 5%), there may be instances where you would accept a larger chance of a Type 1 error to increase power and minimize the chance of a Type 2 error.

**The power/sample size calculation can be expressed in a formula**.

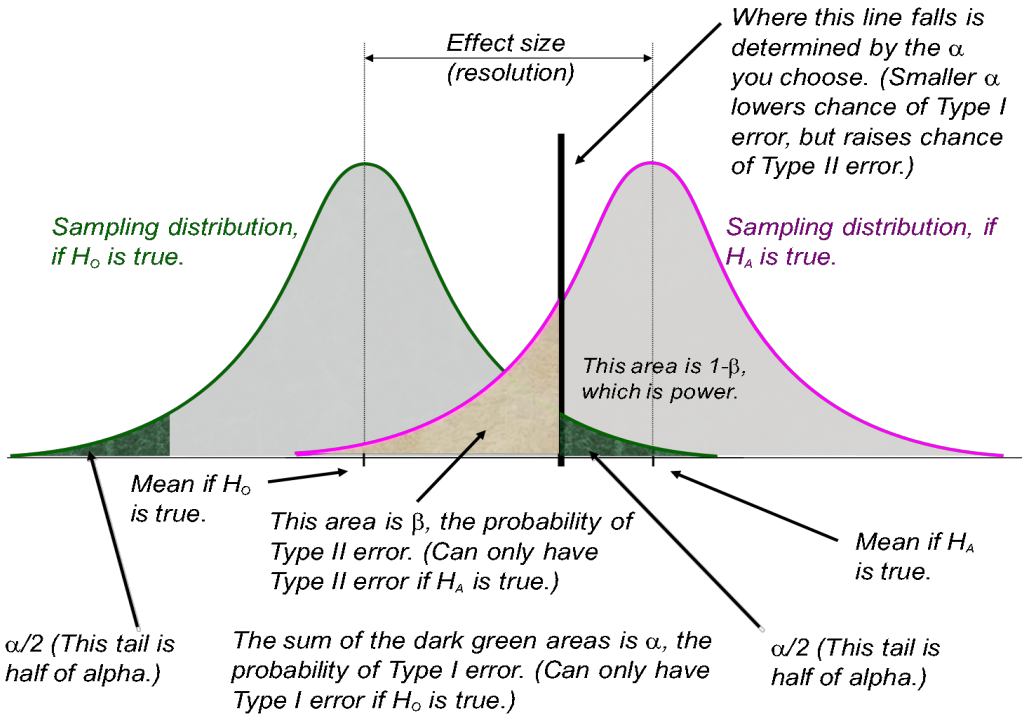
$$n = \frac{2\sigma^2 (Z_{1-\beta} + Z_{1-\alpha/2})^2}{(\mu_1 - \mu_2)^2}$$

A. This formula is for the sample size for a trial comparing two means of a continuous variable, such as blood pressure in patients receiving a drug and placebo, using a t-test
   a. $\mu_1$ and $\mu_2$ = population ("true") means in the two arms.
   b. $\sigma$ (sigma) = standard deviation of the population distribution ($\sigma^2$ is the variance) → a measure of variability
   c. $Z_{1-\beta}$ = value for z (number of standard errors away from zero) corresponding to the power (1-$\beta$). $Z_{1-\beta}$ is a constant for a given power (e.g. for 80% power, $Z_{1-\beta}$ is .84).
   d. $Z_{1-\alpha/2}$ = value for z corresponding to ½ of the risk of pre-specified alpha error. $Z_{1-\alpha/2}$ is 1.96 for a two-sided alpha level of .05.
   e. n = sample size needed **per group.** The total sample size is 2n.
   f. Other formulas exist for sample size calculations for dichotomous and other variables.

You do not need to memorize these formulas, but you do need to understand how power and sample size change with varying levels of alpha, larger or smaller detectable differences, standard deviation, etc.

You should also know that every researcher wants to have a low alpha and a low beta to detect a clinically significant difference. But, the only way to make both lower is to increase sample size or decrease the "noise" (variability) in measurement. For a given sample size and given precision in measurement, there is a tradeoff between alpha and beta (higher power).

**VII. These relationships are also shown graphically in the following figure:**




**VIII. When is it important to think about power?**

**Before a study:**

1. Calculate N (sample size).
2. Or, if N is fixed, to calculate the magnitude of difference that you can detect.

**After the study (or when reading a study):**

1. Interpreting a null result. Is the null hypothesis true or did the study fail to reject the null hypothesis due to inadequate power? Now that you know the magnitude of effect and N, you can determine if the study was adequately powered for that outcome. Also, you can look at the confidence intervals around your effect estimate; if they contain a value would be meaningful (for example, evidence of substantial benefit or harm) then you might want to do a better powered study to confirm/exclude this estimate.
2. Not relevant if a statistically significant result is reported. (If the null hypothesis is rejected, you don't worry about power/Type 2 error, but you still do have to worry about a Type 1 error.)