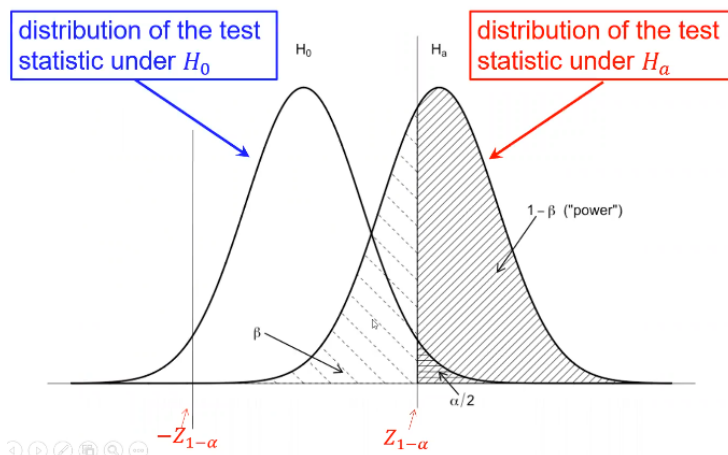


# Power and Sample Size Calculations for Association Studies

Review of errors and difference in means/proportions:

Truth	Your decision	
	Fail to reject $H_0$	Reject $H_0$
<b><math>H_0</math> is True</b>	Correct	Type I error
<b><math>H_0</math> is false</b>	Type II error	Correct

Type II error is represented by beta and type I error as alpha



Where  $Z_{1-(\alpha/2)}$  is the Z-value that creates the target value of alpha in the tail of the distribution. Power decreases as alpha increases.

Power also has a direct relationship with sample size. Sometimes we choose sample sized based on the desired power for a specific alternative hypothesis of interest, and other times we used a fixed sample size and significance level.

A Z-test of differences in proportions with 2 groups with different sample sizes:

$$X_1 = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\left(\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}\right)}}$$

C is the mean of the distribution under the alternative hypothesis:

$$C = \frac{p_1 - p_2}{\sqrt{\left(\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right)}}$$

And the power can be calculated by taking the cdf of C minus the one sided Z-value

$$\text{power} = \Phi(C - Z_{1-\alpha/2}) + \Phi(-C - Z_{1-\alpha/2})$$

Likewise, we can use a similar z-test for differences in the mean of a quantitative trait:

$$X_2 = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}}$$

And the power can be calculated the same way as proportions.

# Applying Power to Genetic Associations

- Difference in proportions: we can compute the power for a test that compares the allele or genotype frequency between cases and controls
- Difference in means: we can compare the mean trait values between individuals with different genotypes at a marker (and we can extend this to an F-test for a linear regression)

## Dichotomous Traits

1. Determine the expected genotype frequencies in cases and controls *under the alternative hypothesis of interest*
2. Estimate power for given sample size or estimate minimum sample size required for given power and significance level *under the planner analysis model*

## Genetic Model

- We usually use 4 parameters to specify the disease model for a dichotomous trait:
  - Genotype relative risks  $y_1$  and  $y_2$
  - Population prevalence of disease K
  - Risk allele frequency  $q$  (the allele that increases risk of disease)
- To compute power, we need the genotype frequencies for cases and controls under the alternative hypothesis
  - $P(\text{person carries } i \text{ risk alleles} \mid \text{affected})$
  - $P(\text{Person carries } i \text{ risk alleles} \mid \text{unaffected})$
  - for  $i = 0, 1, \text{ and } 2$

# Genotype Relative Risk

Another way to parameterize **penetrance**, the probability that someone is affected given their genotype.

Penetrance  $f_i = P(\text{affected} \mid i \text{ risk alleles})$  for  $i = 0, 1$ , or  $2$  risk alleles at a locus

- $f = (f_0, f_1, f_2)$  is the penetrance for the genetic variant.
- $f_0 = f_1 = f_2$  if there is no difference in probability of being affected by genotype  $\rightarrow$  the genotypes at this variant are not associated with case status
- If at least one  $f_i \neq f_j \rightarrow$  the genotype at this variant are associated with case status
- The genotype relative risks (GRR) are ratios of the penetrance values
- $y_i$  is the ratio of probabilities of being a case for someone with  $i$  risk alleles compared to someone with 0 risk alleles
  - $y_1 = f_1/f_0$
  - $y_2 = f_2/f_0$
- $y_1 = y_2 = 1 \rightarrow f_0 = f_1 = f_2 \rightarrow$  No difference in probability of being a case between genotypes with 1 or 2 risk allele(s) and 0 risk alleles  $\rightarrow$  no association between case status and genotype

So if we know the relative risks ( $y_1$  and  $y_2$ ), the population prevalence  $K$ , and the risk allele frequency  $q$ , we can determine the penetrances ( $f_0, f_1, f_2$ ):

$$f_0 = K / (p^2 + 2pq\gamma_1 + q^2\gamma_2) \quad (\text{where } p = 1 - q)$$

and

$$f_1 = f_0\gamma_1$$

$$f_2 = f_0\gamma_2$$

This is derived from the law of total probabilities which I will not show here.

## Models for Dichotomous Traits

| Dominant:  $\gamma_1 = \gamma_2 > 1 \quad \rightarrow (f_0, f_1, f_2) = (f_0, \gamma_2 f_0, \gamma_2 f_0)$

| Recessive:  $\gamma_1 = 1, \gamma_2 > 1 \quad \rightarrow (f_0, f_1, f_2) = (f_0, f_0, \gamma_2 f_0)$

| Multiplicative:  $\gamma_1 = \sqrt{\gamma_2} > 1$

or equivalently  $\gamma_1^2 = \gamma_2 \quad \rightarrow (f_0, f_1, f_2) = (f_0, \sqrt{\gamma_2} f_0, \gamma_2 f_0)$

Note: multiplicative GRR  $\sim$  multiplicative OR

additive  $\log(\text{OR}) \rightarrow$  multiplicative OR

Thus, multiplicative GRR model is most similar to the additive  $\log(\text{OR})$  model that we use in logistic regression

$H_0$ : Genotype or allele frequencies are the same in cases and controls

$H_a$ : genotype frequencies are not equal in cases and controls. They follow a [dominant, recessive, additive] model

To compute power we need the genotype frequencies for the cases and controls under the alternative hypothesis:

$P(i \text{ risk alleles} \mid \text{person is a case})$

$P(i \text{ risk alleles} \mid \text{person is a control})$

For  $i = 0, 1, 2$

The expected genotype frequencies for cases and control from  $f$  using Bayes Rule:

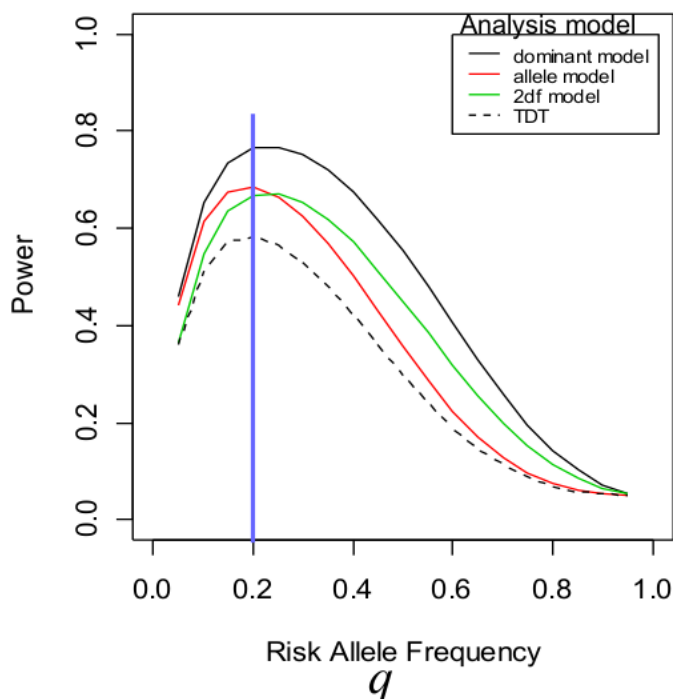
$$\text{Cases: } p_{\text{case}_i} = \begin{cases} p^2 \times f_0 / K & \text{for } i = 0 \\ 2pq \times f_1 / K & \text{for } i = 1 \\ q^2 \times f_2 / K & \text{for } i = 2 \end{cases}$$

$$\text{Controls: } p_{\text{ctrl}_i} = \begin{cases} p^2 \times (1 - f_0) / (1 - K) & \text{for } i = 0 \\ 2pq \times (1 - f_1) / (1 - K) & \text{for } i = 1 \\ q^2 \times (1 - f_2) / (1 - K) & \text{for } i = 2 \end{cases}$$

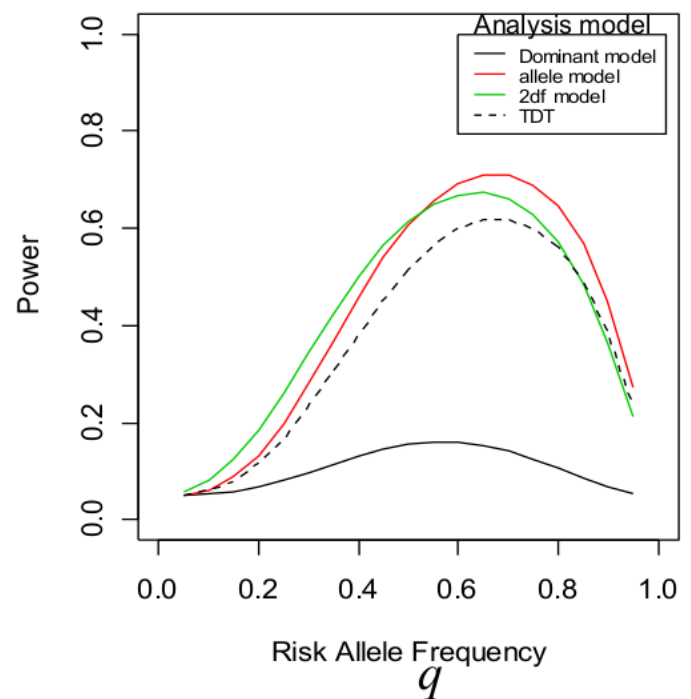
To compute power we not only have to specify the alternative hypothesis model, but also choose the model you will use to do the analysis. The analysis model is the model you choose for the analysis for which you compute power.

Under a dominant model the genotype relative risk should be  $y_1 = y_2 = 2$

100 cases+100 controls or 100 trios (tdt),  
 $H_A$ : **dominant model**, GRR=2



100 cases+100 controls or 100 trios (tdt),  
 $H_A$ : **recessive model**, GRR=2



Notice how when we analyze the recessive model as a dominant one (right), the power decreases significantly.

## Power - General 2DF Test

We can also compute power for the 2df genotype-based chi-squared test. The underlying formulas will not be explained here as there is software to compute this called the Genetic Power Calculator; You specify GRRs, prevalence risk allele frequency and it produces the genotype frequencies in cases and controls and the power under dominant, recessive, additive and 2DF tests.

## Summary - Power for Dichotomous Traits

1. Determine the expected genotype frequencies in cases and controls under the **alternative** hypothesis
  - Need to specify genetic model/effect size
2. Estimate power for given sample size or estimate minimum sample size required for given power using the specific analysis model and significance level
3. Can also find the minimum effect size that yields specific (usually 80%) power for a given sample size and allele frequency

## Quantitative Traits

We usually parameterize power for quantitative traits in terms of locus-specific heritability ( $h^2$ ), the proportion of variance explained by the genetic variant ( $R^2$ ). **It is independent of underlying model and allele frequency.**

For a linear regression F test, the power is determined by the non-centrality parameter  $\lambda$ , the proportion of variance in the trait explained by the genotype ( $h^2$ ) \* sample size

$$\lambda = h^2 * N$$

To determine the (central) F distribution:

In R: `qf(1-alpha, df1, df2)`

Code to determine power using non-central F-distribution and non-centrality parameter "ncp":

In R: `pf(crit.val, df1, df2, ncp, lower.tail = F)`

So for any proportion of variance explained by  $h^2$  and  $N$  we can compute power. However, a single  $h^2$  value corresponds to many alternative hypothesis models.

For dichotomous traits we can simply compute power for difference in proportions, but specifying an underlying genetic model (GRRs, prevalence, allele frequency) is more meaningful. Similarly, power for a quantitative trait is usually more meaningful if we understand how the heritability relates to the genotypic means for a trait -- the genetic model. To find this we need allele

frequency for that allele that increases phenotype and the degree of dominance  $d$  (additive, dominant or recessive).

## Genotypic Means

$a$  is effect size

Recall the variance components in a single gene model:

Genotype:	11	12	22
Genotypic mean ( $g_i$ ):	$-a$	$0$	$a$
Frequency ( $w_i$ ):	$q^2$	$2pq$	$p^2$

Population mean

$$E(g_i) = \bar{\mu} = \sum_{i=1}^3 w_i g_i = p^2 a + 2pqd - q^2 a$$

$$= (p - q)a + 2pqd$$

Locus-specific heritability  $h^2$ , degree of dominance  $d$ , and allele frequency  $p$ :

**Locus-specific heritability:**  $h^2 = \frac{\sigma_G^2}{\sigma_P^2}$

$$\sigma_G^2 = \sigma_A^2 + \sigma_D^2 = \text{Total genetic variance due to locus}$$

For simplicity, let  $\sigma_P^2 = 1$

We can re-scale so that the total trait  $\sigma_P^2 = 1$ , by dividing the trait values by SD:

$$h^2 = \sigma_G^2 = \text{Var}(g_i) = E(g_i - \bar{\mu})^2 = E(g_i^2) - \bar{\mu}^2$$

$$= p^2 a^2 + 2pqd^2 + q^2 a^2 - [(p - q)a + 2pqd]^2$$

⋮ (a bunch of algebra)

$$= 2pq[a - (p - q)d]^2 + 4p^2 q^2 d^2$$

Next we can set  $R = (-1, 0, 1)$  for recessive, additive and dominant models respectively, then  $d = Ra$ . We substitute  $Ra$  for  $d$  in the above and solve for  $a$ :

$$a = \sqrt{\frac{h^2}{2pq[(1 + R(q - p))^2 + 2pqR^2]}}$$

When  $R = 0$ :

$$a = \sqrt{\frac{h^2}{2pq}}$$

## Recap on Effects on Power

- Dichotomous Traits (difference in proportions) are effected by:
  - Effect size: Model (dom/rec/add...)
  - GRRs
  - Risk allele frequency
  - Prevalence of disease/distribution of phenotype
- Quantitative Trait
  - Effect size: Proportion of variance explained by the genetic variant is affected by:
    - Model (dom/rec/add...)
    - Allele frequency

Use a significance level you will use when performing association tests, or the entire family-wise error for the entire experiment.

For the model, if there is no prior hypothesis, try to find examples from the literature of reasonable effect sizes for your alternative hypothesis. Usually under an additive or multiplicative model. Beware that differences in study design can lead to different expected effect sizes.

**Winner's curse** is when significant regression estimates from small underpowered samples are biased -- tend to over-estimate the true effect.

---

Revision #6

Created 8 November 2022 15:49:00 by Elkip

Updated 14 November 2022 13:07:01 by Elkip