Biostatistics 666
Problem Set 5
Due November 27, 2017

Power of Genomewide Association Studies.

Assume that you are evaluating the power of potential genomewide association study where 1,000,000 independent genetic markers will be measured in 2,000 cases of type 2 diabetes and 2,000 population controls.

a) What might be an appropriate significance level for this study?

With ~1,000,000 independent genetic markers to evaluate, an appropriate significance level would be 0.05 / 1,000,000 = cd5x10^{-8}

b) To help evaluate power, fill in the following matrix:

Power as a function of Allele Frequency and Effect Size

	Relative Risk		
Population Allele Frequency (f)	Modest (r=1.1)	Moderate (r=1.3)	Large (r=1.5)
Low (f = 0.05)	0%	0.8%	26.9%
Moderate (f = 0.20)	0%	49.2%	99.9%
Common (f = 0.50)	0.1%	85.1%	100%

***Assuming a multiplicative model and assuming the population prevalence of type 2 diabetes is 10%.**

To fill in the matrix, I recommend you use a spreadsheet or a little bit of R code (it would be tedious to repeat calculations for each cell otherwise). In Excel, the normal distribution functions NORMSDIST() – the cumulative normal distribution – and NORMSINV() – the inverse of cumulative normal distribution function; will come in handy. In R, the equivalent functions are pnorm() and qnorm().

Briefly, note that N = N_{cases} = $N_{controls}$ = 2,000, that prevalence K = 10%, that relative risk r = 1.1, 1.3 or 1.5 (depending on column) and that allele frequency f = 0.05, 0.20 or 0.50 (depending on row). Also, note that significance level $\alpha = 5\text{x}10^{-8}$ (corresponding to C = 5.33, since $\Phi^{-1}(5\text{x}10^{-8}) = -5.33$).

To calculate power, we first calculate:

$$p_{case} = \frac{p(1-p)r + p^2r^2}{(1-p)^2 + p(1-p)r + p^2r^2}$$

And, then:

$$p_{control} = \frac{p - p_{case}K}{1-K}$$

Which allows us to calculate the expected tested statistics:

$$E(Z) = \mu = \frac{p_{case} - p_{control}}{\sqrt{[p_{case}(1 - p_{case}) + p\_control(1 - p_{control})]/2N}}$$

Then, to calculate power for each cell, we evaluate:

$$Power = P(Z < -C) + P(Z > C) = \Phi(-5.33 - \mu) + 1 - \Phi(5.33 - \mu)$$

c) What sample size would be required to achieve 80% power for a low frequency allele (f = 0.05) that makes a modest contribution to disease risk (r = 1.1)?

On the order of 79,200 cases and 79,200 controls would be required.

d) If your budget is limited and genotyping the number samples suggested in c) is outside your budget, how might you increase power for detecting alleles that make modest contributions to disease risk?

There are a number of strategies for increasing power for a given cost:

- Look for opportunities to use external controls, which have already been genotyped
- Use a smaller genotype array, followed by imputation, to reduce cost
- Execute a two stage-design where only a fraction of cases is genotyped for all variants
- Focus on a set of cases that are enriched for genetic causes of disease (they have an affected relative or are younger or leaner than expected in the case of type 2 diabetes, for example).

Affected Sibling Pair Analysis

In a study of type 2 diabetes (population prevalence = 5%), investigators collected 10 affected sibling pairs. These pairs were genotyped for a polymorphism with two alleles, allele "1" with frequency 0.20 and allele "2" with frequency 0.80. Genotyping results show that for all 10 pairs one sibling has genotype "2/2" and the other sibling has genotype "1/1".

a) **If a genetic variant in the locus being examined had a population frequency of 10% and increased the risk of type 2 diabetes by 2-fold for heterozygotes and 4-fold for homozygotes, what would be the expected IBD proportions among affected sibling pairs?**

So, we have an allele with frequency p_- = 0.10 which results in genotype frequencies of $p_{-/-}$ = 0.01, $p_{+/-}$ = 0.18 and $p_{+/+}$ = 0.81. We know the risk of disease for these three genotypes is 4*f*, 2*f* and *f*.

Then, we have:

$$K = 4pf + 4p(1-p)f + (1-p)^2 f = 1.21f$$

(although we don't know *f*, we know it must be <0.25 so that the probability of disease for a risk allele homozygote, which is 4*f*, is <1.)

Then,

$$\lambda_{MZ} = \frac{16p^2f^2+8p(1-p)f^2+(1-p)^2f^2}{K^2} = \frac{1.69f^2}{1{,}46f^2} = 1.16$$

$$\lambda_O = \frac{16p^3f^2+16p^2(1-p)f^2+4p(1-p)f^2+4p(1-p)^2f^2+(1-p)^3f^2}{K^2} = \frac{1.57f^2}{1.46f^2} = 1.07$$

$$\lambda_S = 0.25 + 0.50\,\lambda_O + 0.25\,\lambda_{MZ} = 1.08$$

Finally,

$$P(IBD = 0|ASP) = 0.25\frac{1}{\lambda_S} = 0.23$$

$$P(IBD = 1|ASP) = 0.50\frac{\lambda_O}{\lambda_S} = 0.50$$

$$P(IBD = 2|ASP) = 0.25\frac{\lambda_{MZ}}{\lambda_S} = 0.27$$

So, in this case, we can see that the sharing proportions are not very far from those under the null, and we might well expect relatively low power for an affected sibling pair linkage test.

b) **Calculate the LOD score for this dataset using the MLS test of Risch (1990). Does the result suggest there is evidence for linkage?**

One thing to note here: because the siblings pairs all have different genotypes, with no alleles in common, we know they must all be IBD=0. Therefore …

$$\hat{z}_0 = 1, \hat{z}_1 = 0, \hat{z}_2 = 0$$

$$LOD = \log_{10} \prod_i \frac{z_0}{0.25} = 6$$

Although the LOD score is high, we should be cautious with interpretation. We know that for a true genetic linkage signal, the probability of sharing 2 alleles IBD must be >0.25 and the probability of sharing 0 alleles IBD must be <0.25.

c) **Would the possible triangle constraint of Holmans (1993) affect the estimated LOD score? Why is the constraint useful?**

With the possible triangle constraint, the LOD score would be zero. The constraint is useful because it restricts analysis to the set of models compatible with a genetic association signal.

d) **Differences between reported and actual relationships among the individuals being studied can affect the power of a genetic linkage study. Speculate how this might be important.**

Genetic linkage studies compare actual and expected patterns of IBD sharing. When relationships are misspecified, we will use incorrect expectations for sharing. For example, if two individuals are described as siblings, we will expect them to share a chromosome IBD 50% of the time and look for sharing in excess of 50%. If they are actually half-siblings, null sharing will be 25% and maximum sharing will be 50%, so tests for sharing beyond 50% will lose power. Similarly, if the two are actually identical twins, null sharing will be 100%, and a test assuming they are full-siblings would have excess type 1 error.

e) **Genotyping error is another common challenge for genetic studies. Speculate how this might be important.**

Genotyping error is a challenge because it makes estimation of IBD more challenging. Since genotyping error is almost always a possibility, it is important for analysis methods to account for this possibility when modeling IBD sharing patterns.