GWAS 3: Statistical power

Matti Pirinen, University of Helsinki

Updated: March 12, 2025

The slide set referred to in this document is "GWAS 3".

We saw previously that a stringent significance threshold, such as 5e-8, is needed in GWAS in order to avoid false positives, that is, such null variants that reach the significance threshold. On the other hand, a stringent threshold makes it quite difficult to get even the true non-zero effects to reach the threshold. In other words, we tend to have many false negatives, that is, non-zero variants that do not reach the significance threshold.

Next, we will study the properties of the variants and the study design that determine how likely we are to catch the true effects. This topic is called statistical power analysis (Slides 1-7). Here is a review article on statistical power in GWAS by Sham and Purcell 2014.

Statistical power of a statistical significance test is the probability that the test will reject the null hypothesis H_0 (in GWAS, H_0 says that $\beta = 0$) at the given significance threshold when the data follow a specific *alternative hypothesis* H_1. In the GWAS setting, H_1 is specified by fixing the study design (total sample size or case and controls counts) and the parameters defining the variant (MAF and effect size).

To compute the P-value, we only needed to consider the null hypothesis H_0. For a power analysis, we also need to define explicitly how the true effects look like, i.e., we need to quantify the alternative hypothesis H_1. Of course, not all true effects are the same, and therefore power analysis is often presented as a power curve over a plausible range of parameter values.

3.1 Test statistic under the alternative By assuming that the sampling distribution of the effect size estimate $\widehat{\beta}$ is Normal (which works well for large sample sizes and common variants), we have that $\widehat{\beta} \sim \mathcal{N}(\beta, \text{SE}^2)$, where β is the true effect size and SE is the standard error of the estimator. As we saw previously, under the null hypothesis H_0, i.e. when $\beta = 0$, the Wald test statistic $z = \widehat{\beta}/\text{SE}$ is distributed approximately as $z \sim \mathcal{N}(0, 1)$, which can be used for computing a P-value. Another way to compute the same P-value is via the chi-square distribution as $z^2 \sim \chi^2_1$. With the chi-square distribution, we need to consider only the upper tail of the distribution to compute the two-sided P-value, whereas with the Normal distribution we would need to remember to multiply by 2 the tail probability to get the same two-sided P-value.

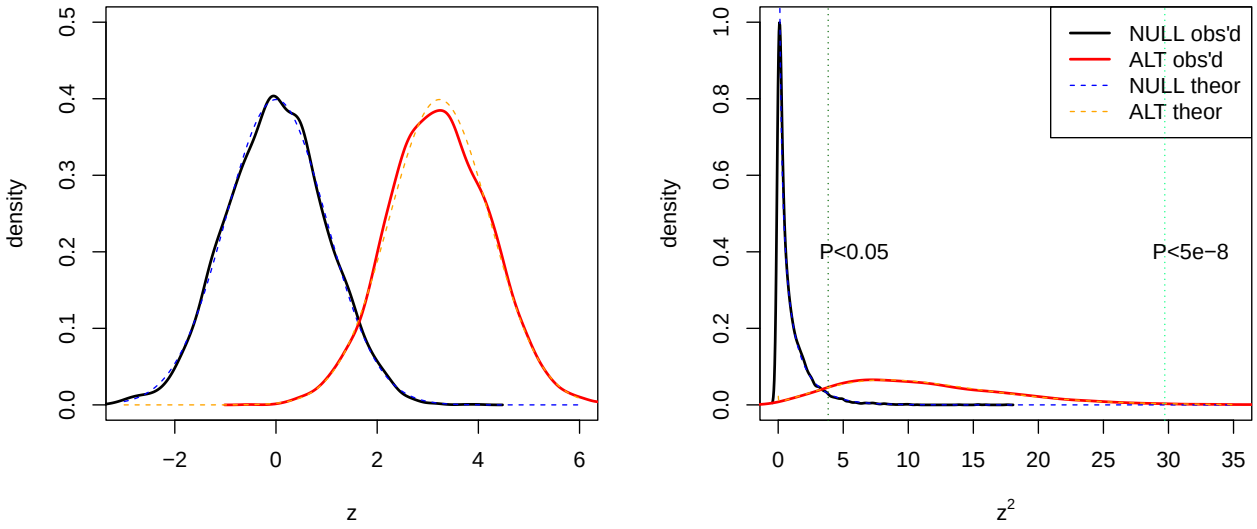
What happens when $\beta \neq 0$, that is, when the variant has a non-zero effect? Then $z \sim \mathcal{N}(\beta/\text{SE}, 1)$ and $z^2 \sim \chi^2_1((\beta/\text{SE})^2)$, which is called a chi-square distribution with 1 degree of freedom and non-centrality parameter NCP=$(\beta/\text{SE})^2$. When $\beta = 0$ this reduces to the standard *central* chi-square distribution $\chi^2_1 = \chi^2_1(0)$.

Example 3.1. Let's illustrate these distributions by a simulation of GWAS results under both the alternative hypothesis and under the null hypothesis. To save time, we don't do regressions with genotype-phenotype data but we simulate the effect estimates directly from their known distributions. First, however, we need to find the standard error, and that we do by fitting one regression model. We will visualize the distributions both for the Wald statistic (having a Normal distribution) and for its square (having a chi-square distribution).

```r
n = 500 # individuals
p = 5000 # SNPs for both null and alternative
f = 0.5 # MAF
b.alt = 0.2 # effect size under the alternative hypothesis
x = rbinom(n, 2, f) # genotypes at 1 SNP for n ind
y = scale( rnorm(n) ) # random phenotype normalized to have sample sd=1
se = summary( lm( y ~ x ) )$coeff[2,2] # pick SE, and assume it stays constant and independent of beta
b.hat.null = rnorm(p, 0, se) # estimates under null
b.hat.alt = rnorm(p, b.alt, se)  # estimates under alternative

par(mfrow=c(1,2))
# Plot observed densities of z-scores
plot(NULL, xlim = c(-3,6), ylim = c(0,0.5), xlab = "z",
     ylab = "density", col = "white") # empty panel for plotting
lines(density( (b.hat.null/se) ), col = "black", lwd = 2) # Wald statistic for null variants
lines(density( (b.hat.alt/se) ), col = "red", lwd = 2) # Wald statistic for alternative variants
# add theoretical densities for z-scores
x.seq = seq(-3, 6, 0.01)
lines(x.seq, dnorm(x.seq, 0, 1), col = "blue", lty = 2) # for null
lines(x.seq, dnorm(x.seq, b.alt/se, 1), col = "orange", lty = 2) # for alternative

# Plot observed densities of z^2
plot(NULL, xlim = c(0,35), ylim = c(0,1), xlab = expression(z^2),
     ylab = "density", col = "white") # empty panel for plotting
lines(density( (b.hat.null/se)^2 ), col = "black", lwd = 2) # chi-square stat for null variants
lines(density( (b.hat.alt/se)^2 ), col = "red", lwd = 2) # chi-square stat for alternative variants
# Let's add theoretical densities of the chi-square distributions
x.seq = seq(0, 35, 0.01)
lines(x.seq, dchisq(x.seq, df = 1, ncp = 0), col = "blue", lty = 2) # ncp=0 for null
lines(x.seq, dchisq(x.seq, df = 1, ncp = (b.alt/se)^2), col = "orange", lty = 2) # ncp = (beta/se)^2 fo
legend("topright", leg = c("NULL obs'd","ALT obs'd","NULL theor","ALT theor"),
       col = c("black","red","blue","orange"),
       lty = c(1,1,2,2), lwd = c(2,2,1,1) )
# Let's add significance thresholds corresponding to 0.05 and 5e-8
# By definition, the thresholds are always computed under the null.
q.thresh = qchisq( c(0.05, 5e-8), df = 1, ncp = 0, lower = FALSE)
abline(v = q.thresh, col = c("darkgreen", "springgreen"), lty = 3)
text( q.thresh+2, c(0.4,0.4), c("P<0.05","P<5e-8") )
```



The theoretical distributions match well with the observed ones so we trust that we now understand the relevant parameters also of the theoretical chi-square distribution. We also see that almost the whole of the alternative distribution is to the right of the significance threhsold of 0.05 but to the left of threshold 5e-8. Thus, with these parameters, we would discover almost all variants at a significance level 0.05 but almost none at the genome-wide significance level of 5e-8.

How to compute the exact proportion of the distribution to the right of a given threshold value? We first compute the cut points of the test statistics under the null hypothesis using `qchisq()` function and then we compute the upper tail probabilities at these cut points for the non-central chi-square distribution corresponding to the alternative hypothesis.

```
q.thresh = qchisq(c(0.05,5e-8), df = 1, ncp = 0, lower = FALSE) # signif. thresholds in chi-square unit
pchisq(q.thresh, df = 1, ncp = (b.alt/se)^2, lower = FALSE) # corresponding upper tail probabilities
```

```
## [1] 0.89821524 0.01321279
```

So we have a probability of 90% to detect a variant at significance threshold 0.05, when effect size is 0.2 SD units of a quantitative phenotype, MAF is 50% and the sample size is 500. This probability is also called statistical **power** corresponding to the given parameters (significance threshold, β, MAF and n). On the other hand, with these parameters, we only have a power of 1.3% at the genome-wide significance level 5e-8. An interpretation is that we are likely to discover 90 out of 100 variants having the parameters considered at the significance level 0.05 but only about 1 out of 100 at the level 5e-8.

A lower bound for power is the significance threshold α. We use the concept of statistical power to describe the ability to detect *non-zero effects*. What about the zero effects? What is the probability of getting a significant result when the null hypothesis holds? By definition, this probability is the significance threshold α, and does not depend on n or MAF. Consequently, the power to detect any non-zero effect can never be less than α and will get close to α for tiny effects ($\beta \approx 0$) which are almost indistinguishable from 0.

Typically, we would like to design studies that have a large power (say $\geq 80\%$) to detect the types of variants that we are interested in. How can we do that?

3.2 Ingredients of power

The parameters affecting power are

1. Sample size n; increasing sample size increases power because it increases accuracy of effect size estimate.

2. Effect size β; increasing the absolute value of effect size increases power because it increases difference from the null model.

3. Minor allele frequency f; increasing MAF increases power because it increases accuracy of effect size estimate.

4. Significance threshold α; increasing the threshold increases power because larger significance thresholds are easier to achieve.

5. In case-control studies, the proportion of cases ϕ; moving ϕ closer to 0.5 increases power because it increases accuracy of effect size estimate.

We will soon discuss intuition why each of these parameters affects power. But let's first write down how these parameters and power are tied together.

For a given significance level α, power is determined by the non-centrality parameter NCP$= (\beta/\text{SE})^2$ of the chi-square distribution. The mean of the distribution is 1+NCP and the whole distribution moves to right with increasing NCP. Hence, the larger the NCP the larger the power. We see that the NCP increases with β^2 and this explains why increasing $|\beta|$ increases power. We also see that the NCP increases as SE decreases and therefore we need to know how SE depends on n, f and ϕ.

3.2.1 Formulas for standard errors For the linear model

$$y = \mu + x\beta + \varepsilon,$$

SE of $\widehat{\beta}$ is

$$\text{SE}_{\text{lin}}\left(\widehat{\beta}\right) = \frac{\sigma}{\sqrt{n\text{Var}(x)}} \approx \frac{\sigma}{\sqrt{2nf(1-f)}},$$

where the variance of genotype x is, under Hardy-Weinberg equilibrium, approximately $2f(1-f)$, and σ is the standard deviation of the error term ε: $\sigma^2 = \text{Var}(y) - \beta^2\text{Var}(x)$. This form for SE is a direct consequence of the variance estimate of the mean-centered linear model: $\text{Var}(\widehat{\beta}) = \sigma^2/\sum_{i=1}^n x_i^2$.

In a typical quantitative trait GWAS, the effects of variants on the total phenotypic variance are small ($<$ 1%) and then we can assume that the error variance $\sigma^2 \approx \text{Var}(y)$, which is approximately equal to 1, if the phenotype is defined via quantile normalization or via scaling of the residuals after regressing out the covariate effects.

For *binary* case-control GWAS analyzed by logistic regression,

$$\text{SE}_{\text{bin}}\left(\widehat{\beta}\right) \approx \frac{1}{\sqrt{n\text{Var}(x)\phi(1-\phi)}} \approx \frac{1}{\sqrt{2nf(1-f)\phi(1-\phi)}}.$$

Thus, the difference from the linear model SE is that σ is replaced by 1 and n is replaced by an **effective sample size** $n\phi(1-\phi)$. Here n is the total sample size, i.e., cases + controls. For derivation, see Appendix A of Vukcevic et al. 2012. Note: Often the effective sample size is defined as $4n\phi(1-\phi)$, because that quantity tells what would be the total sample size (cases + controls) in a hypothetical study that has equal number of cases and controls and whose power matches the power of our current study.

A smaller SE means a higher precision of the effect size estimate. Both of the formulas show how SE decreases with increasing sample size n, with increasing MAF f and, for binary data, SE decreases as ϕ gets closer to 0.5. These formulas work well for typical GWAS settings but may not hold when some parameter (n or f or $\phi(1-\phi)$) gets close to zero. In particular, the formula may not be good for the rare variant testing ($f < 0.001$). To know exactly when the formulas start to break down, it is best to do simulations.

Now we can write down the NCPs of the additive GWAS models as

$$\text{NCP}_{\text{lin}} = (\beta/\text{SE}_{\text{lin}})^2 \approx 2f(1-f)n\beta^2/\sigma^2 \qquad \text{and} \qquad \text{NCP}_{\text{bin}} = (\beta/\text{SE}_{\text{bin}})^2 \approx 2f(1-f)n\phi(1-\phi)\beta^2.$$

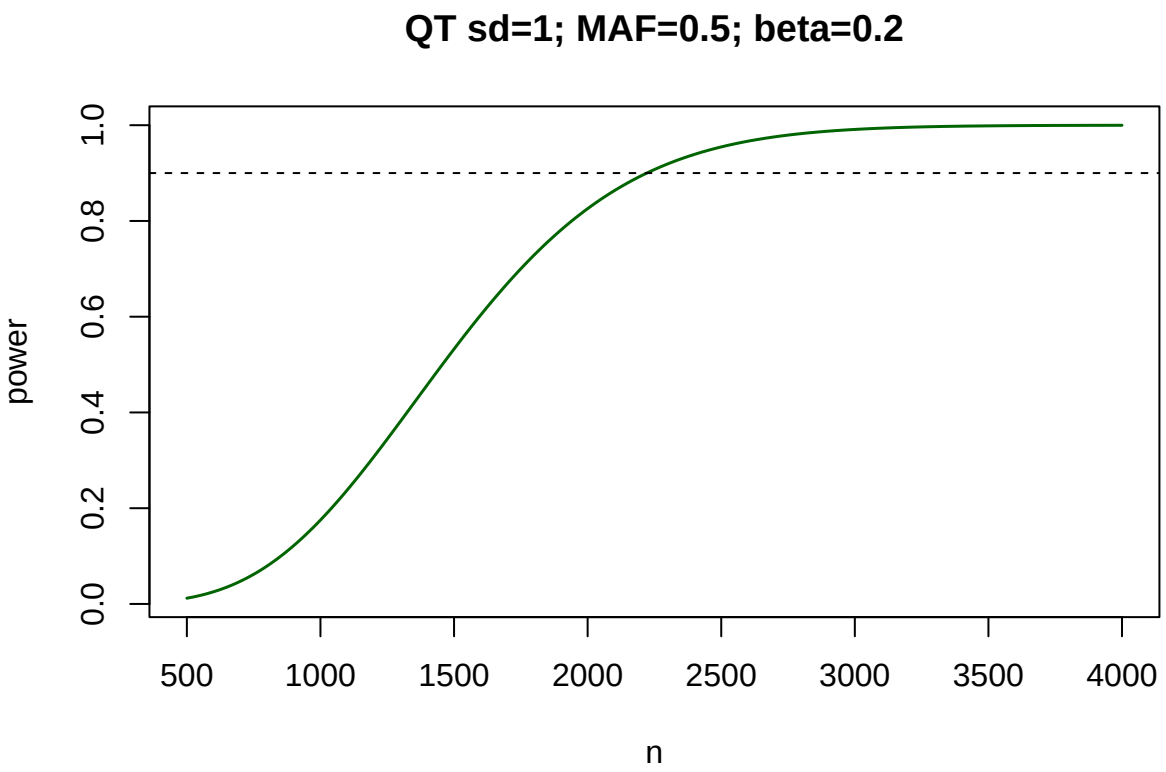
Let's next discuss how and why each parameter affects the NCPs, and hence the power.

3.2.2 Sample size Out of the parameters affecting power, the sample size is most directly under the control of study design. Therefore, it is the primary parameter by which we can design studies of sufficient power.

Increasing n decreases SE in the regression models in proportion to $1/\sqrt{n}$. In the GWAS context, we can think that a larger n leads to more accurate estimate of the phenotypic means in each of the genotype classes. Therefore, as n grows, we also have more accurate estimate of the phenotypic difference between the genotype classes, which means that we are better able to distinguish a true phenotypic difference between the genotype groups. In other words, we have a larger power to detect a genotype's effect on the phenotype.

Example 3.2. Above we saw that with $n = 500$ (and MAF $= 0.5$, $\beta = 0.2$) we had only 1% power at significance level $\alpha =$ 5e-8. Let's determine how large n should be to achieve 90% power.

```
f = 0.5
b.alt = 0.2
sigma = sqrt(1 - 2*f*(1 - f)*b.alt^2) # error sd after SNP effect is accounted for (see next part for e
ns = seq(500, 4000, 10) # candidate values for n
ses = sigma/sqrt(ns*2*f*(1 - f)) # SEs corresponding to each candidate n
q.thresh = qchisq(5e-8, df = 1, ncp = 0, lower = FALSE) # chi-sqr threshold corresponding to alpha = 5e
pwr = pchisq(q.thresh, df = 1, ncp = (b.alt/ses)^2, lower = FALSE) # power at alpha = 5e-8 for vector o
plot(ns, pwr, col = "darkgreen", xlab = "n", ylab = "power",
     main = paste0("QT sd=1; MAF=",f,"; beta=",b.alt), t = "l", lwd = 1.5)
abline(h = 0.9, lty = 2)
```



```r
# Let's output the first n that gives power >= 90%
ns[min(which(pwr >= 0.9))]
```

```
## [1] 2230
```

So, we need $n = 2230$ in order to have power of 90%.

3.2.3 Effect size and variance explained When the effect size is β and the MAF is f, then the variance explained by the additive effect on the genotype is $\text{Var}(x\beta) = \text{Var}(x)\beta^2 \approx 2f(1-f)\beta^2$. When the total phenotypic variance of a quantitative trait is 1, then $2f(1-f)\beta^2$ is also the proportion of the variance explained by the variant. For example, in our ongoing example setting, the variance explained by the variant is

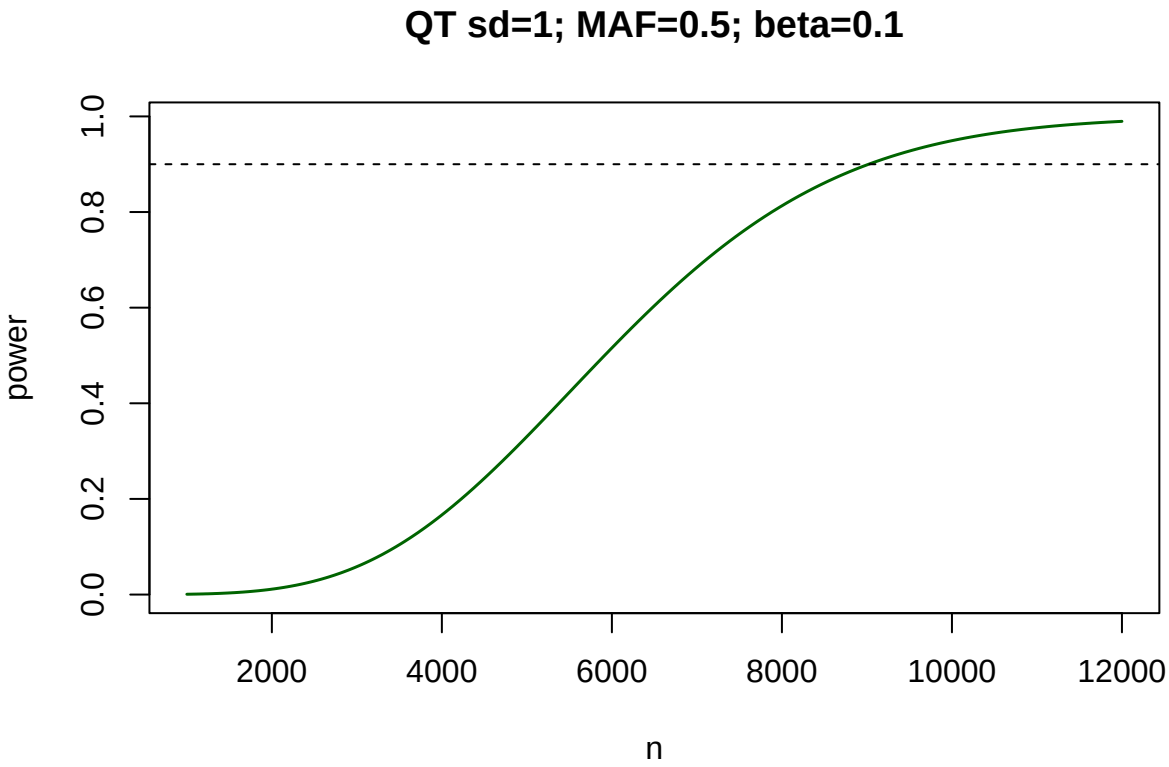
```r
2*f*(1-f)*b.alt^2.
```

```
## [1] 0.02
```

That is, the variant explains 2% of the variation of the phenotype. This is a very large variance explained compared to a typical common variant association with complex traits, such as BMI or height, but is more realistic for some molecular traits, such as metabolite levels, that are less complex genetically and may be affected by larger effects from individual variants.

Example 3.3. What if we wanted to find a suitable n that gives 90% power for MAF=50% when the variant explained only 0.5% of the phenotype?

```r
f = 0.5
y.explained = 0.005
b.alt = sqrt(y.explained / (2*f*(1 - f)) ) # this is beta that explains 0.5%
sigma = sqrt(1 - y.explained) # error sd after SNP effect is accounted for
ns = seq(1000, 12000, 10) # candidate n
ses = sigma / sqrt( ns*2*f*(1 - f) ) # SE corresponding to each n
q.thresh = qchisq(5e-8, df = 1, ncp = 0, lower = FALSE) # threshold corresp. alpha = 5e-8
pwr = pchisq(q.thresh, df = 1, ncp = (b.alt/ses)^2, lower = FALSE) # power at alpha = 5e-8
plot(ns,pwr, col = "darkgreen", xlab = "n", ylab = "power",
     main = paste0("QT sd=1; MAF=",f,"; beta=",b.alt), t = "l", lwd = 1.5)
abline( h = 0.9, lty = 2 )
```




```
# Let's output n that is the first that gives power >= 90%
ns[min(which(pwr >= 0.9))]
```

```
## [1] 9030
```

So, we needed to multiply the sample size by a factor of ~4. This really makes a difference in practice. It is very different to collect 2230 individuals than to collect 9030. Power calculations are important!

Could we maybe had guessed the factor of 4 without doing the actual power calculation? For a fixed α, power is determined by the NCP. Earlier we determined the parameters that gave 90% power. If we equate the NCP defined by those parameters with an NCP cooresponding to a new effect size and an unknown sample size n_2, we can solve for n_2. Furthermore, as $f = 0.5$ remains constant, it cancels out, and we get

$$\frac{2f(1-f)\beta_1^2 n_1}{\sigma_1^2} = \frac{2f(1-f)\beta_2^2 n_2}{\sigma_2^2} \longrightarrow n_2 = \frac{\beta_1^2 n_1 \sigma_2^2}{\beta_2^2 \sigma_1^2} = n_1 \frac{0.2^2}{0.1^2} \frac{1-0.005}{1-0.02} \approx 4.0612 \cdot n_1.$$

We conclude that when the variance explained by the variant is small (say $< 2\%$), and we drop the variance explained by a factor of a, (which is the same as dropping the effect size by a factor of $\sqrt{a}$), we must increase the sample size approximately by a factor of a to maintain constant power. If variance explained by the variant is larger than a few percents, then it will have a bigger effect on the result, but in GWAS, typically, the variance explained remains $< 1\%$.

3.2.4 Minor allele frequency When the other parameters than the MAF f remain constant, the NCPs are proportional to $f(1-f)$ which is maximized at $f = 0.5$ and which decreases to zero as $f \to 0$.

Technically, this can be explained by a general property that, in regression models, the increased variance in the predictor variables makes the effect estimates more precise. Considering more concretely the GWAS setting, if MAF f is very small, then almost all individuals in the sample have the major homozygote genotype. Consequently, we will estimate the mean phenotype very accurately for that homozygote genotype,

but we will have almost no information about the phenotypic mean of the heterozygote group. Therefore, we have little information whether the two groups are *different* from each other. The mathematics show that, from the point of view of power, the best balance between the precision of the phenotypic mean estimates of different genotype groups is achieved when MAF=0.5.

Example 3.4. Consider a situation where the MAF of a particular variant is 0.5 in population A but only 0.2 in population B. How much larger sample size would we need in B to achieve the same power as in A assuming all other parameters were the same?

Answer. Let's equate the NCPs and, since the other parameters than f and n cancel out, we are left with

$$n_B f_B(1-f_B) = n_A f_A(1-f_A) \longrightarrow n_B = n_A \frac{f_A(1-f_A)}{f_B(1-f_B)} = n_A \frac{0.5 \times 0.5}{0.2 \times 0.8} = 1.5625 \cdot n_A.$$

So we need 56% more samples in B than in A.

We can get a simpler approximation for cases where MAF is low (<5%) in both populations, because then $(1-f_A)/(1-f_B) \approx 1$ and we are left with $n_B \approx f_A/f_B\, n_A$.

Example 3.5. In section 1, we encountered *PCSK9* gene's missense variant rs11591147 that had a strong effect on the LDL-cholesterol levels. GnomAD database shows that it has MAF 4.2% (713/16994) in Finland whereas MAF is 1.5% (1235/81394) in non-Finnish Europeans (NFE). Hence, we estimate that we would need to collect about $4.2/1.5 = 2.8$ times larger cohort in NFE than in FIN to detect this variant at a given significance threshold. (The exact factor is $0.042 \cdot (1-0.042)/(0.015 \cdot (1-0.015)) = 2.723$.)

3.2.5 Proportion of cases In a case-control analysis of a binary trait, the proportion of cases in the sample, ϕ, affects power. Mathematically, the effect is similar to that of MAF, i.e., the NCP is proportional to $\phi(1-\phi)$. Hence, all other things being fixed (including the total sample size), the largest power is achieved when the number of cases equals the number of controls ($\phi = 0.5$), and power approaches zero when $\phi \to 0$ or $\phi \to 1$. Here, intuition is that if we had only few cases, then the allele frequency information in cases were very inaccurate and therefore it would be very difficult to determine whether case and control frequencies are *different* from each other, no matter how accurately we could estimate the control allele frequency.

Example 3.6. Consider two GWAS on migraine.

1. The UK Biobank study of 500,000 individuals of whom 15,000 have self-reported migraine.
2. Case-control analysis of 60,000 individuals of whom 30,000 suffer from migraine and 30,000 are controls. Which of these two studies yields larger power?

Answer. The power is determined by the NCP=$2f(1-f)n\phi(1-\phi)\beta^2$, and the two studies are assumed to differ only in their effective sample size $n\phi(1-\phi)$. Therefore, the one that has the larger effective sample size has also larger power.

```
n = c(500000, 60000)
phi = c(15000, 30000)/n
cbind(n, phi, eff.n = n*phi*(1 - phi))
```

```
##          n  phi eff.n
## [1,] 5e+05 0.03 14550
## [2,] 6e+04 0.50 15000
```

These studies are very similarly powered but the second study has a slightly higher power. This is true even though the total sample size of the second study is only 12% of the total sample size of the first study.

3.2.6 Power calculators Genetic Association Study Power Calculator at University of Michigan specifies case-control study effect sizes by genotype relative risk (GRR) and disease prevalence in population. For their multiplicative model, GRR is the relative risk between genotype 1 and 0 as well as between genotype 2 and 1. It can be shown that for low prevalence (<1%) GRR approximately equals to the odds ratio (OR). Let's test it.

Example 3.7. Make a study with 10,000 cases and 10,000 controls, with disease allele frequency 40%, GRR of 1.1 and disease prevalence of 0.001, in which case GRR is approximately OR. The calculator gives power 22.5% at $\alpha = 5e\text{-}8$. Compare it to the way we have done it:

```
b = log(1.1) #b is log-odds, approximately GRR for a low prevalence disease
n = 20000
f = 0.4
phi = 0.5
pchisq(qchisq(5e-8, df = 1, lower = F), df = 1, ncp = 2*f*(1-f)*n*phi*(1-phi)*b^2, lower = FALSE)
```

```
## [1] 0.2170825
```

Methods agree well and estimate power of 22..23%.

3.2.7 Risk allele frequency (RAF) vs minor allele frequency (MAF) Continue with the Power calculator and set GRR to 1.15 and RAF to 0.2: power is 60%. Change RAF to 0.8; power drops to 51%. How can that be since in our formula the power depends only on MAF which is the same (0.2) in both cases?

Let's consider a disease with low prevalence (say < 1%). Then the RAF in disease-free controls is almost the same as RAF in the whole population since over 99% of the population are eligible to become controls. By definition, the risk allele is more frequent among the cases than among the controls and hence the RAF in the ascertained case-control sample, that is heavily enriched for cases compared to their population frequency, will be higher than RAF in the population. Hence, if the risk allele is also the minor allele, then the MAF in the ascertained case-control sample will be higher than in the population, which increases the power. However, if the risk allele is the major allele, then the MAF in the ascertained case-control sample will be lower than in the population, which decreases the power. We conclude that in an ascertained case-control sample, we have more power to detect risk increasing minor alleles than protective minor alleles even when their MAFs were the same in population Chan et al. 2014 AJHG. This reminds us about the importance of considering f in the power calculations as the MAF in the particular sample that we are analysing, which for ascertained case-control studies may differ from the population's MAF.

3.3 Why well powered studies are so important?

The GWAS evolution over the last 15 years gives an illuminating lesson on the influence of statistical power on scientific conclusions. Since the effect sizes of common variants affecting complex diseases are relatively small, the first years of GWAS of any one disease were able to detect only few GWS associations because the sample sizes were only a few thousands. However, when the sample size grew to tens of thousands, the number of associations started a steep increase. This pattern has occurred for nearly all traits and diseases studied. (Demonstrated for schizophrenia on slides 8-10.)

Example 3.8. Let's look at the results of the Schizophrenia study ("Biological insights from 108 schizophrenia-associated genetic loci")[https://www.nature.com/articles/nature13595], Nature 511:421-427 taken from their Supplementary table 2.

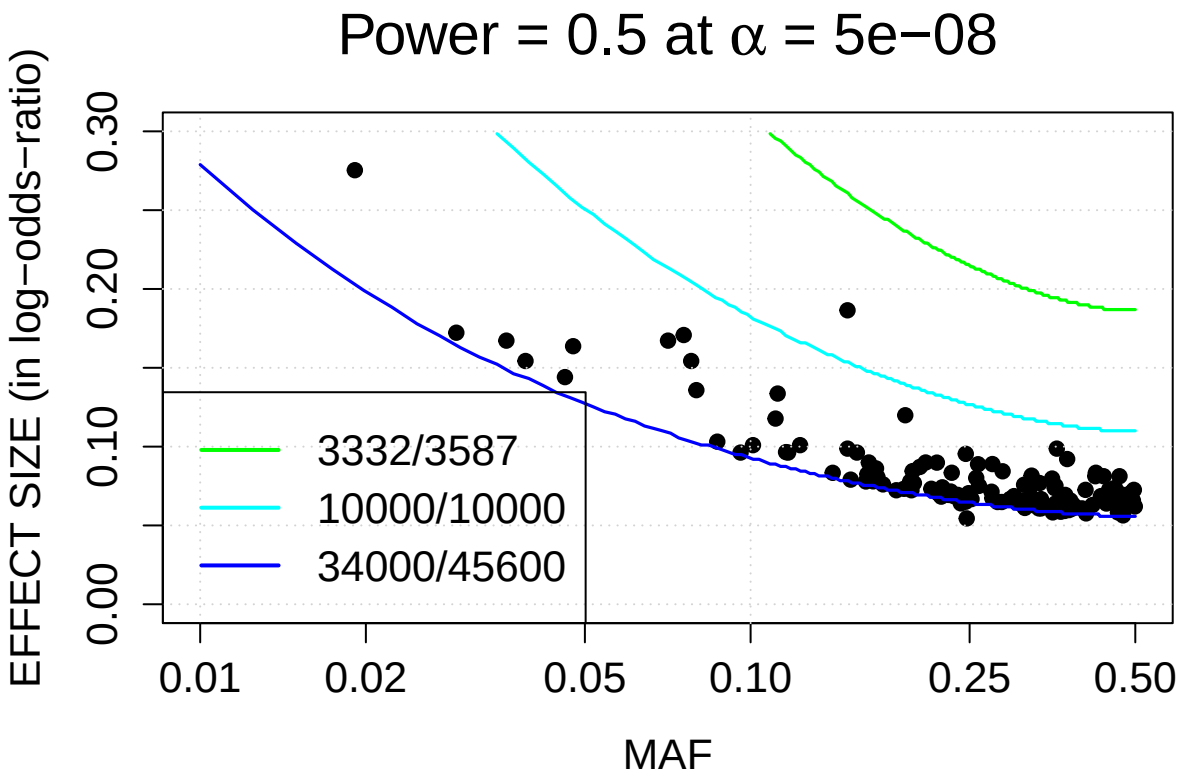
```
sz.res = read.table("http://www.mv.helsinki.fi/home/mjxpirin/GWAS_course/material/sz_res.txt",
                    as.is = TRUE, header = TRUE)
sz.res[1,] #see what data we have
```

```
##   Index_SNP A12 Frq_case Frq_control Chr Start_Position End_Position
## 1 rs4648845  TC    0.533       0.527   1        2372401      2402501
##   Combined_OR Combined_95lower Combined_95upper Combined_P Discovery_OR
## 1       1.072            1.049            1.097    8.7e-10        1.071
##   Discovery_P Replication_OR Replication_P
## 1    4.03e-09          1.088        0.0885
```

```
#Let's plot the known SZ variants on frequency - effect size coordinates
#And draw some power curves there at genome-wide significance threshold
maf = sz.res[,"Frq_control"] #Not yet maf but allele 1 frequency
maf[maf > 0.5] = 1 - maf[maf > 0.5] #Make it to MAF: always less than 0.5
b = abs(log(sz.res[,"Combined_OR"])) #effect size on log-odds-ratio scale with positive sign
pw.thresh = 0.5
p.threshold = 5e-8
plot(maf, b, ylim = c(0, 0.3), xlim = c(0.01, 0.5), xlab = "MAF",
     ylab = "EFFECT SIZE (in log-odds-ratio)", xaxt = "n", yaxt = "n", log = "x", #make x-axis logarith
     main = substitute(paste("Power = ", pw.thresh ," at ", alpha ," = ",p.threshold),
                       list(pw.thresh = pw.thresh, p.threshold = p.threshold)),
     cex.main = 1.8, cex.lab = 1.3, pch = 19)
axis(1, at = c(0.01, 0.02, 0.05, 0.10, 0.25, 0.5), cex.axis = 1.3)
axis(2, at = c(0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3), cex.axis = 1.3)
grid()

q = qchisq(p.threshold, df = 1, lower = FALSE) #chi-square value corresp. significance threshold
#matrix of numbers of cases (col1) and controls (col2):
Ns = matrix( c(3332,3587,
               10000,10000,
               34000,45600),
             ncol = 2 , byrow = TRUE)
cols=c("green", "cyan", "blue")

f = seq(0.01, 0.5, length = 200)
b = seq(0, 0.3, length = 200)
legends = c()
par(mar = c(6, 6, 5, 1))
for(set in 1:nrow(Ns)){
  pw = rep(NA, length(b)) #power at each candidate b
  b.for.f = rep(NA,length(f)) #for each f gives the b value that leads to target power
  for(i in 1:length(f)){
    pw = pchisq(q, df = 1, ncp = Ns[set,1]*Ns[set,2] / sum(Ns[set,])*2*f[i]*(1 - f[i])*b^2, lower = FALS
    b.for.f[i] = b[ min( which(pw > pw.thresh) ) ]
  }
  lines(f, b.for.f, t = "l", col = cols[set], lwd = 1.6)
  legends = c(legends, paste(Ns[set,],collapse = "/") ) #make a "#cases/#controls" tag for legend
}
legend("bottomleft", lty = c(1, 1), col = cols, legend = legends, lwd = 2, cex = 1.3)
```



No wonder that the 2009 study with 3300 cases and 3600 controls (slide 8) did not find any individual SNPs associated with SZ because it had insufficient power for any now know (common) SZ variant. Note that none of these GWS variants has MAF < 0.02, most likely because rare variants have not yet been comprehensively analyzed in these large meta-analyses.

3.3.1 Absence of evidence is not evidence of absence It is important to understand that a "non-significant" P-value should not be interpreted as evidence that there is no effect. It can be so interpreted only for those effect sizes for which the power to detect them was ~100%. But a "non-significant" P-value does not rule out smaller effects. Therefore, any claims that argue based on statistical evidence that an effect does not exist must be made precise by stating which power there would had been to detect effects, as a function of quantities of interest, such as β and MAF. Additionally, the effect size estimate and its SE should be reported in addition to the "non-significant" P-value.

Example 3.9. Suppose that we study migraine and find a variant v that has an association P-value 3.2e-15 in a large migraine GWAS, a P-value of 2.1e-9 in a subset of cases with *migraine without aura* and a P-value of 0.08 in a subset of cases with *migraine with aura*. Would you interpret this as evidence that the variant v is associated with migraine, and, in particular, that this effect is specific to migraine without aura and is not present in migraine with aura? What other information would you need to make this conclusion? Describe some example study setting (by describing sample sizes and effect sizes) where the above conclusion would be appropriate, and another study setting where it would not be appropriate.

3.3.2 Proportion of true positives Let's recall the formula for the odds of a significant P-value indicating a true positive rather than a false positive:

$$\frac{P(T \mid S)}{P(N \mid S)} = \frac{P(T)P(S \mid T)}{P(N)P(S \mid N)} = \text{prior-odds} \times \frac{\text{power}}{\text{significance threshold}}.$$

Thus, for a fixed significance threshold and prior-odds of association, the probability of a significant result being a true effect increases proportionally to the power of the study. Hence, a larger proportion of the

significant findings from a well powered study is likely to be true positives than from an underpowered study. Another way to think this is that all studies (independent of their power) have the same rate of labelling null effects as significant (and this rate is α, the significance level, the Type I error rate) but different rates of labelling true effects as significant (and this rate is the power of the study). Hence, by increasing power, a larger proportion of all significant results will be true positives.

3.3.3 Winner's curse Suppose that we have 50% power to detect MAF=5% variants with effect $\beta = 0.2$ with our GWAS sample at GWS threshold. This means that, in our data, half of the variants whose true effect is 0.2 will reach the threshold and other half does not. How are these two groups different in our data? They are different in that those that reach the threshold have estimated $\widehat{\beta} > 0.2$ whereas those that don't reach the threshold have estimated $\widehat{\beta} < 0.2$. Consequently, out of all those variants with true $\beta = 0.2$ and MAF=5%, the ones that become GWS in our data have their effect sizes **overestimated**. This is the *winner's curse*: when power is low enough, the variants can reach the GWS threshold only if their effect sizes are overestimated. We are "winners" because we are able to detect some of these true effects but we are simultaneously "cursed" because we have upwardly biased effect size estimates. And this curse gets worse as the power decreases. Originally, the term winner's curse relates to auctions where the winner is the one who has made the biggest offer and the biggest offer tends to be higher than the consensus value of the item; the winner is always cursed to pay too much.

Note that the winner's curse, as defined above, does not occur for variants that have so large effect sizes that power to detect them is ~100%. For such variants, we still overestimate the effect size in 50% of variants and underestimate it in the remaining 50% of variants, but now we will detect essentially all of those variants at the given significance threshold, even those whose effects we underestimate. Hence, with high power, there is no general trend of the significant variants' effects being overestimated. In practice this means that we are less concern with the winner's curse for GWS variants that have P-values, say < 1e-20, than for those that have P-values just slightly below 5e-8.

Of note, we may induce another flavor of winner's curse when we choose the variant with the smallest P-value from a GWAS locus to represent the locus. The logic is that if, in a GWAS locus, we have many correlated variants that truly have similar effect sizes, then the particular one that happens to have the smallest P-value in our data set tends rather to have an overestimated effect size than an underestimated effect size in our data set.

How to get rid of winner's curse? There are some statistical methods for that (of course!) but the simplest way is to get an effect size estimate from an independent replication data set. Since those replication data were not used for the original GWS finding, there is no reason to expect that the effect size estimate would be (upwardly) biased also in the replication data.

When deciding on the size of the replication study, it is good to remember that, due to the winner's curse in the original study, the raw effect size estimate from the discovery GWAS (i.e., the GWAS that made the GWS discovery in the first place) may result in a too low sample size estimate if used in the power calculation.

3.3.4 Reasonable study design If a study proposal has low power to answer the question of interest, it is difficult to argue why that study should be funded. The funding agencies want to know that the studies they fund will have a reasonable chance to provide solid conclusions.

Power calculations are crucial in medical studies that study treatment options on patients or animals because there either too small or too large cohorts are unethical. Too small cohorts sacrifice some individuals by exposing them to potentially harmful treatments without ever producing solid results, whereas too large cohorts will expose unnecessarily many individuals to a suboptimal treatment, even after available statistical evidence could had already told which is the optimal treatment.

In GWAS studies, it is typical that each new GWAS data collection will be combined with the already existing ones, and hence the joint power of all existing data is the scientifically most relevant quantity to consider.