Modeling Genetic Traits

Randy Johnson

4/13/2017

Setup

```r
library(Rtools)
library(GWASTools)
library(SNPRelate)
library(tidyverse)
library(cowplot)
library(broom)
library(ggplot2)
library(stringr)

theme_update(text = element_text(size = 20))

# colorblind palette
cbbPalette <- c("#000000", "#E69F00", "#56B4E9", "#009E73",
                "#F0E442", "#0072B2", "#D55E00", "#CC79A7")
```

Sample data set

We aren't going to analyze these exact data, but this is what the each dataset will look like.

```
## # A tibble: 400 × 3
##             cont   cat     a
##            <dbl> <int> <int>
## 1  -0.802527689     1     1
## 2  -0.939746757     0     2
## 3  -0.004529563     1     1
## 4   1.144833439     1     1
## 5   1.139871137     1     0
## 6   1.290047785     0     2
## 7   1.691404159     1     2
## 8  -0.312656551     1     2
## 9  -0.031561397     1     0
## 10  0.799614626     1     1
## # ... with 390 more rows
```

Additive traits

- Definition: The effect of each phenotype influencing variant changes the phenotype by an equal ammount for each inherited allele. Example: human skin color. It is unknown how many genes affect human skin color, but additivity is fairly well established.

```
glm(cat ~ (a == 1) + (a == 2), data = dat, family = binomia
    tidy() %>%
    mutate(OR = exp(estimate)) %>%
    select(term, OR, p.value)
```

```
##          term       OR     p.value
## 1 (Intercept) 1.088608 0.585900962
## 2  a == 1TRUE 1.968439 0.002575053
## 3  a == 2TRUE 2.694574 0.003282868
```

Multiplicitive traits

- Definition: The effect of each phenotype influencing variant changes the phenotype by a constant multiplier. For example, if one allele increases gene expression by 2 fold, two alleles will increase gene expression by 4 fold. Example: Hemoglobin A/S. Individuals with with a heterozygous hemoglobin phenotype (i.e. they have one A gene and one S gene) will experience very slight symptoms similar to sickle cell anemia. Individuals with two hemoglobin S genes will present with sickle cell disease. The effect of having two alleles is worse than double the symptoms in heterozygous individuals.

```
glm(cat ~ a, data = dat, family = binomial) %>%
    tidy() %>%
    mutate(OR = exp(estimate)) %>%
    select(term, OR, p.value)
```

```
##          term        OR      p.value
## 1 (Intercept) 0.9825317 9.108590e-01
## 2           a 2.3903625 2.008948e-07
```

Multiplicitive traits

This is often called an additive model within the context of logistic regression, because at the log Odds scale, it is additive. At the OR scale, however, it is multiplicitive. Be sure to be very clear when defining your terms in the methods section.

- log Odds scale

$$\begin{aligned}\log(OR_1|a=1) &= \beta_1\\ \log(OR_1|a=2) &= 2*\beta_1\\ &= \beta_1+\beta_1\end{aligned}$$

- Odds Ratio (OR) scale

$$\begin{aligned}(OR_1|a=1) &= e^{\beta_1}\\ (OR_1|a=2) &= e^{\beta_1+\beta_1}\\ &= e^{\beta_1}*e^{\beta_1}\end{aligned}$$

Dominant traits

- Definition: The phenotype is observed if there are one or two variants. Example: Polydactyly (extra fingers and toes) is a dominant trait caused by one of a number of different genes (e.g. GLI3). The allele frequency is about 2%.

```r
glm(cat ~ (a > 0), data = dat, family = binomial) %>%
    tidy() %>%
    mutate(OR = exp(estimate)) %>%
    select(term, OR, p.value)
```

```
##          term        OR      p.value
## 1 (Intercept) 0.9324324 6.759210e-01
## 2   a > 0TRUE 2.6025121 9.883431e-06
```

Recessive traits

- Definition: The phenotype is observed only if there are two variant alleles. Example: CCR5 Δ 32 homozygosity provides near perfect protection against HIV infection.

```r
glm(cat ~ (a == 2), data = dat, family = binomial) %>%
    tidy() %>%
    mutate(OR = exp(estimate)) %>%
    select(term, OR, p.value)
```

```
##          term        OR    p.value
## 1 (Intercept) 0.9248555 0.47633196
## 2   a == 2TRUE 2.2116477 0.00493205
```

X-linked traits

- Definition: Genes with recessive traits that are found on the X chromosome are X-linked. Males have only one X chromosome, so the presence of a recessive gene will not be compensated for by another chromosome inherited from the father. Because the gene is located on the X chromosome, males are more likely to be affected. Example: Hemophilia. Females with only one defective FVIII or FIX gene will still produce functional clotting factors in the blood stream, and the phenotype is not observed.

```
glm(cat ~ dummy, family = binomial,
    data = mutate(dat, dummy = a == 2 |
                                  (a == 1 & male))) %>%
    tidy() %>%
    mutate(OR = exp(estimate)) %>%
    select(term, OR, p.value)
```

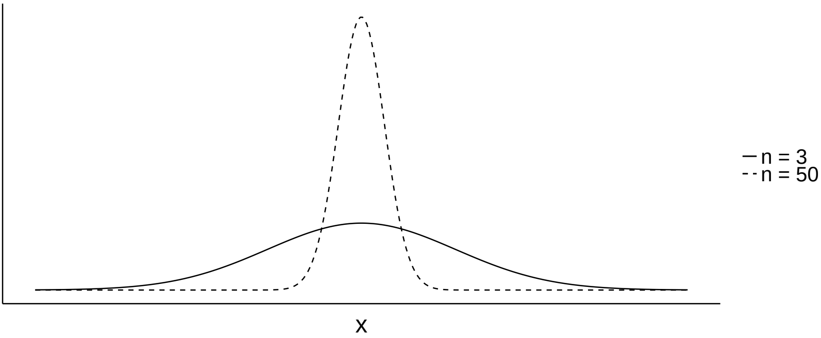
```
##          term       OR      p.value
## 1 (Intercept) 1.058394 6.338399e-01
## 2   dummyTRUE 2.651613 5.100375e-05
```

Asside: Mosaic traits

- Definition: This is another type of X-linked trait, affecting females. In females, one X chromosome is essentially deactivated and forms what is called a Barr body. This happens relatively early in the development of the unborn offspring, and clonal expansions of daughter cells, all with the same X chromosome Barr body, will result in patches of similar X chromosome characteristics. Thus a gene affecting phenotype may be observable in patches. Example: Orange/black coat color in cats. Tortiseshell and Calico cats are always female.

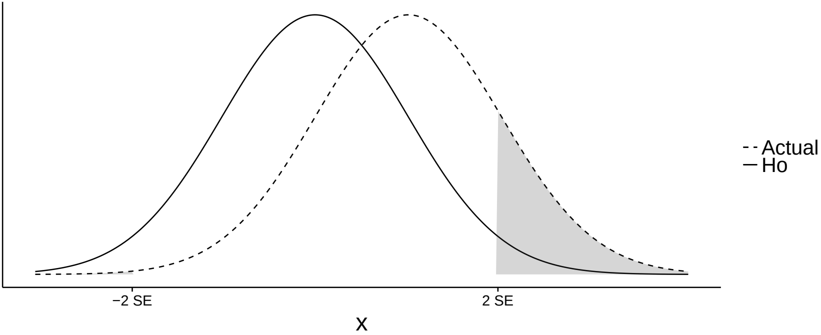
Background: Power

- By the Central Limit Theorem (CLT), the mean of your observed test distribution will be closer to the actual population mean as the sample size increases.
- By the Law of Large Numbers (LLN), the variance of your test distribution will get smaller as the sample size increases.



Background: Power

The concept of statistical power comes into play when there is a null hypothesis to compare our test statistic against.

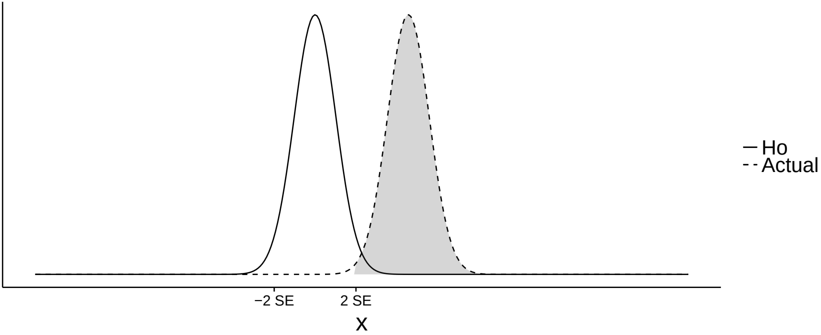


Given these sampling distributions, what is the probability that we correctly reject the null hypothesis? This is statistical power.

Background: Power and Sample Size

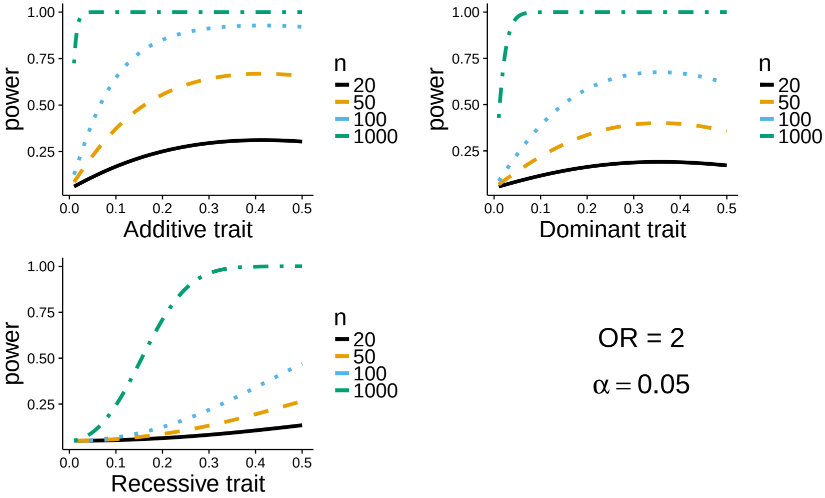
Now, we can increase the probability of correctly rejecting the null hypothesis if we increase the sample size.

- Why? Because the CLT and LLN are our friends!



Power of Genetic Models

So, how much power do we have for different genetic models? How often will we be able to reject the null hypothesis when we should?



OR = 2

$\alpha = 0.05$

Creation of GDS files from PLINK output

```
# Create binary plink files
#  looks for: ASW.ped and ASW.map
#  creates: 'plink.bed', 'plink.bim', 'plink.fam'
system('plink --file ASW --make-bed --noweb')

# R package will expect gzipped files
system('gzip plink.*')

# Convert to GDS
snpgdsBED2GDS('plink.bed.gz', 'plink.fam.gz',
              'plink.bim.gz', 'ASW.gds',
              family=TRUE, cvt.chr="int",
              cvt.snpid="int", verbose=FALSE)
```

Merging GDS files

```
snpgdsCombineGeno(paste0('~/Documents/HapMap3/gds/',
                         c('CHB', 'CHD'), '.gds'),
                  '~/Documents/HapMap3/gds/CH.gds')

## Create ~/Documents/HapMap3/gds/CH.gds
##        with 169 samples and 1258957 SNPs
##    Open the gds file ~/Documents/HapMap3/gds/CHB.gds.
##        0 strands of SNP loci need to be switched.
##    Open the gds file ~/Documents/HapMap3/gds/CHD.gds.
##        46997 strands of SNP loci need to be switched.
```

Opening GDS files

```
(gds <- GdsGenotypeReader('~/Documents/HapMap3/gds/CH.gds'))

## File: /Users/johnsonra/Documents/HapMap3/gds/CH.gds (59.6M)
## +    [  ] *
## |--+ sample.id   { Str8 169 ZIP_ra(27.7%), 381B }
## |--+ snp.id   { Str8 1258957 ZIP_ra(34.9%), 4.3M }
## |--+ snp.position   { Int32 1258957 ZIP_ra(81.7%), 3.9M }
## |--+ snp.chromosome   { Int32 1258957 ZIP_ra(0.10%), 4.9K }
## |--+ snp.allele   { Str8 1258957 ZIP_ra(14.1%), 692.8K }
## \--+ genotype   { Bit2 1258957x169, 50.7M } *
```

Adding Sample Annotations

```
gds <- GdsGenotypeReader('~/Documents/HapMap3/gds/CH.gds')
# Be sure clinical data agrees with GDS file
load('../project1/dat4.RData')

dat4 <- filter(dat4, IID %in% getVariable(gds, 'sample.id')) %>%
        rename(scanID = IID) %>%
        right_join(data_frame(scanID =
                                      getVariable(gds, 'sample.id'))) %>
        as.data.frame()
```

```
## Joining, by = "scanID"
```

```
close(gds)
```

Run Logistic Regression

```
geno <- GenotypeData(data = gds,
                     scanAnnot = ScanAnnotationDataFrame(dat4))

# get chromosome number (only analyze autosomes here)
chr <- getVariable(gds, 'snp.chromosome')

out <- assocRegression(geno, 'hiv', model.type = 'logistic',
                       gene.action = 'additive',
                       covar = c('multiple_partners',
                                 'share_needles',
                                 'protected_sex'),
                       snpStart = 1,
                       snpEnd = max(which(chr < 23)))

## Reading in Phenotype and Covariate Data...
## Running analysis with 84 Samples
## Beginning Calculations...
## Block 1 of 245 Completed - 19.4 secs
## Block 2 of 245 Completed - 14.24 secs
## Block 3 of 245 Completed - 14.91 secs
##
```

Logistic Regression Output

```r
load('../Data/15_logisticOut1.RData')
str(out)
```

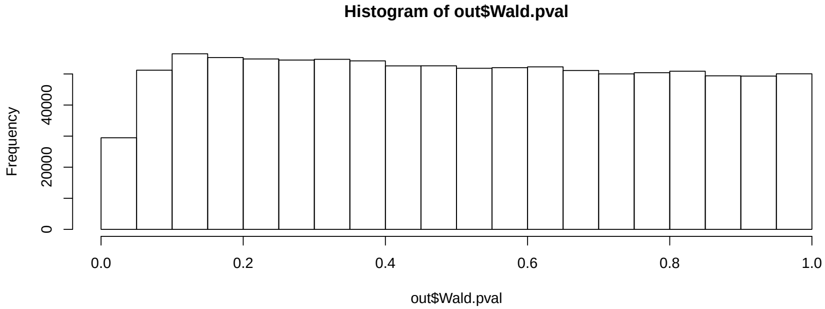
```
## 'data.frame':    1220353 obs. of  14 variables:
##  $ snpID        : chr  "rs4124251" "rs6650104" "rs10458597" "
##  $ chr          : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ effect.allele: chr  "A" "A" "A" "A" ...
##  $ EAF          : num  0.2857 0.0427 0.0488 0.0122 0.0952 ...
##  $ MAF          : num  0.2857 0.0427 0.0488 0.0122 0.0952 ...
##  $ n            : num  84 82 82 82 84 84 84 82 81 84 ...
##  $ n0           : num  7 7 7 7 7 7 7 7 7 7 ...
##  $ n1           : num  77 75 75 75 77 77 77 75 74 77 ...
##  $ Est          : num  0.451 -0.332 0.124 NA -0.577 ...
##  $ SE           : num  0.893 1.353 0.986 NA 1.046 ...
##  $ LL           : num  -1.3 -2.98 -1.81 NA -2.63 ...
##  $ UL           : num  2.2 2.32 2.06 NA 1.47 ...
##  $ Wald.Stat    : num  0.2545 0.0601 0.0157 NA 0.3039 ...
##  $ Wald.pval    : num  0.614 0.806 0.9 NA 0.581 ...
```

Logistic Regression Output

```
summary(-log10(out$Wald.pval))
```

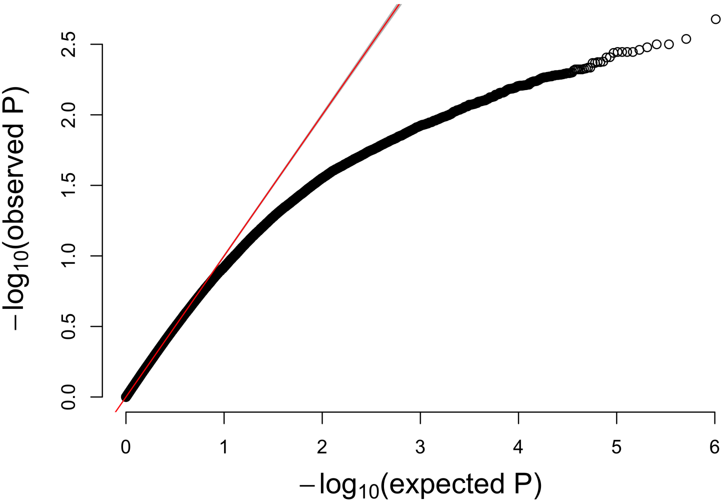
```
##     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##     0.00    0.13    0.30    0.41    0.59    2.68  197256
```

```
hist(out$Wald.pval)
```



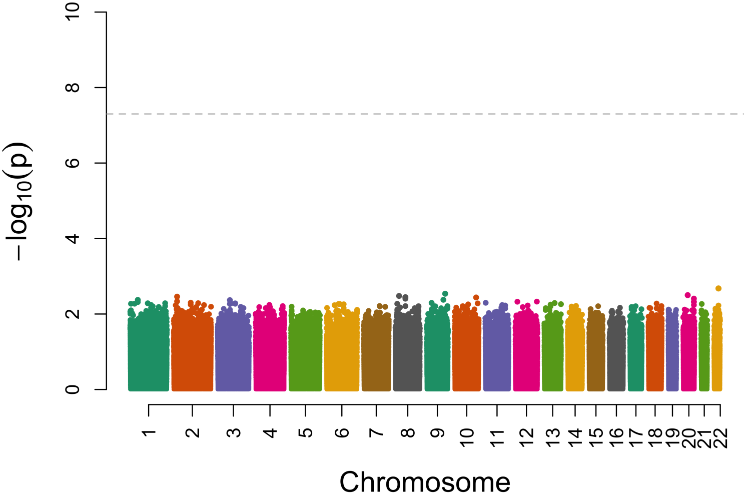
Logistic Regression Output

```
qqPlot(out$Wald.pval)
```



Logistic Regression Output

```
manhattanPlot(out$Wald.pval, out$chr)
```



Logistic Regression Output

```r
arrange(out, Wald.pval)[1:10,] %>%
    select(snpID, chr, Est, Wald.pval)
```

```
##         snpID chr       Est   Wald.pval
## 1    rs932514  22 -3.680179 0.002100185
## 2   rs4836817   9 -3.745232 0.002898329
## 3   rs4814934  20 -3.296442 0.003164243
## 4   rs6035539  20 -3.296442 0.003164243
## 5   rs7003304   8 -3.849518 0.003322244
## 6  rs12710685   2 -3.538091 0.003463502
## 7   rs1477958   8 -4.465193 0.003585230
## 8   rs4873755   8 -4.465193 0.003585230
## 9   rs1477953   8 -4.465193 0.003585230
## 10  rs4873754   8 -4.465193 0.003585230
```

Run Survival Analysis

```
gds <- GdsGenotypeReader('~/Documents/HapMap3/gds/CH.gds')
geno <- GenotypeData(data = gds,
            scanAnnot = {mutate(dat4,
                          thiv = ifelse(thiv > 0, thiv, NA)) %>%
                          ScanAnnotationDataFrame()})

out2 <- assocCoxPH(geno, event = 'hiv',
                   time.to.event = 'thiv',
                   gene.action = 'additive',
                   covar = c('multiple_partners',
                             'share_needles',
                             'protected_sex'),
                   snpStart = 1, snpEnd = 100)
```

```
## Reading in Phenotype and Covariate Data...

## Running analysis with 49 Samples

## Beginning Calculations...
```

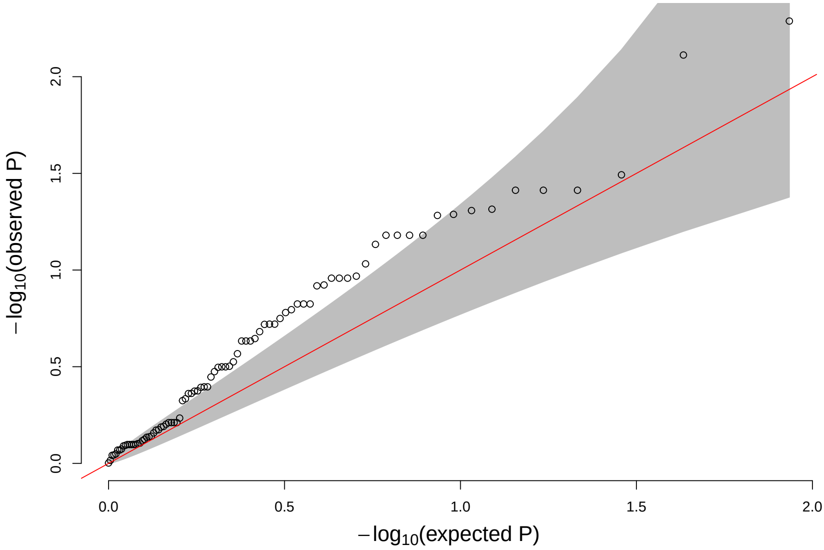
Survival Analysis Output

We only ran the first 100 here, but you will use the output techniques above for survival analysis as well (note that we will use `z.pval`, though).

```
##         snpID chr        Est      z.pval
## 1   rs9442398   1  1.3599045 0.005154518
## 2   rs9442387   1 -1.2163296 0.007724840
## 3   rs3737728   1  1.1294830 0.032174172
## 4  rs17160824   1  0.4791745 0.038666155
## 5  rs11807848   1  0.4791745 0.038666155
## 6  rs12757754   1  0.4791745 0.038666155
## 7  rs13303118   1  0.6335336 0.048455529
## 8   rs2298217   1 -0.3975030 0.049245351
## 9   rs9442373   1 -0.3923876 0.051533138
## 10 rs10907182   1 -0.4010702 0.052186403
```

Survival Analysis Output

This only has 100 data points, but for the full plot you probably want to save it to a png.



Significance Threshold

If I pulled a coin out and flipped 10 heads in a row, would you be impressed?

```r
set.seed(6107)

rbinom(n = 10, size = 1, prob = 0.5)
```

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

```r
# probability:
0.5^10
```

```
## [1] 0.0009765625
```

Fishing Experiment

What really happend:

```
set.seed(9382)
randomSeeds <- round(runif(1000) * 10000)

for(s in randomSeeds)
{
    set.seed(s)
    heads <- sum(rbinom(n=10, size=1, prob=0.5))
    if(heads == 10)
        break
}
s
## [1] 6107
```

Type I and Type II Errors

Probability of making a Type I error (i.e. incorrectly rejecting the null hypothesis when it is true. . . i.e. calling something significant when it isn't):

$$1 - (1 - \alpha)^n$$

when there are *n* tests in your study.

Type I and Type II Errors

Applying this formula to our regression results, the probability of a Type I error is:

α	Error Rate
0.05	$>$0.999
1×10^{-5}	$>$0.999
1×10^{-6}	0.705
1×10^{-7}	0.115
5×10^{-8}	0.059
1×10^{-8}	0.012

Bonferroni Corrections

A simple approximation to put a p-value into perspective with the family-wise Type I error rate is:

$$p_{corrected} = p * n$$

or

$$\alpha_{study\ \ wide} = \frac{\alpha}{n}.$$

So, our Bonferroni-corrected significance threshold would be

$$\frac{0.05}{1,220,353} \approx 4.1 \times 10^{-8}$$

Declaring Significance

People often get hung up on what the significance threshold should be for their study. A Bonferroni correction is probably too conservative in most instances, but what threshold should you use?

If your results are close enough to the Bonferroni significance threshold that you are wondering if you can squeeze a little extra statistical power out of your analysis, then they probably aren't that exciting to begin with. They very well might be publishable, though! Share what you have found, *and let others know exactly what you did.*

What can you do if you lack power?

Power can be thought of as the probability that you will be able to reject the null hypothesis when you should. Not rejecting the null hypothesis when you should is called a Type II error.

So, what can you do if you lack statistical power? What can you do when you have little hope of being able to reject the null hypothesis when the alternate hypothesis is true?

- Give up and move onto something else.
- Collect more data (design your studies to avoid this).
- Publish what you have an hope someone else (perhaps you) will be able to replicate your results in another study.