Modeling Genetic Traits

Randy Johnson

4/13/2017

Setup

```r
library(tidyverse)
library(cowplot)
library(broom)
library(ggplot2)

theme_update(text = element_text(size = 20))

# colorbline 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.

```
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.

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

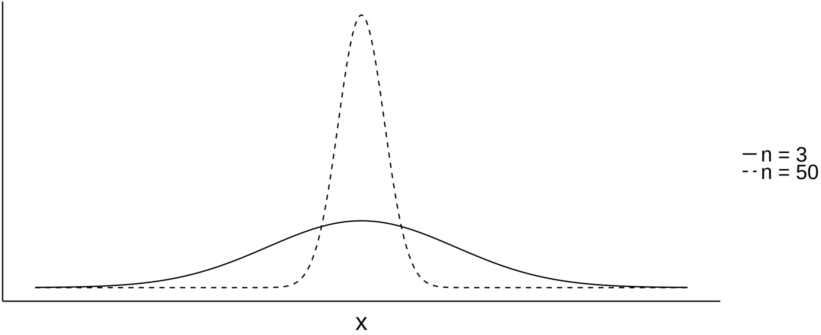
```
##          term       OR      p.value
## 1 (Intercept) 1.058394 6.338399e-01
## 2   dummyTRUE 2.651613 5.100375e-05
```

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.
- Power when you are right/wrong
- Power under different minor allele frequencies

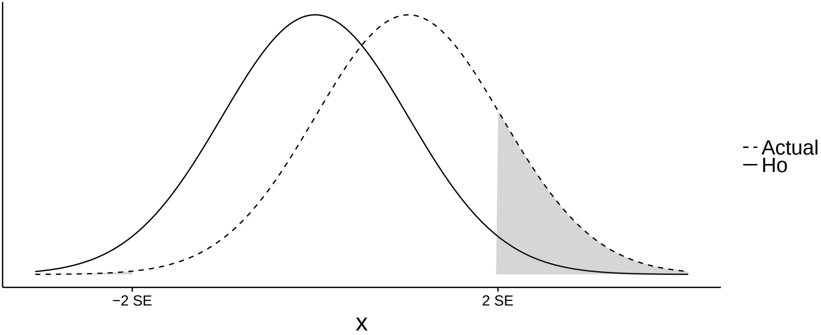
Background: Power

- 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.

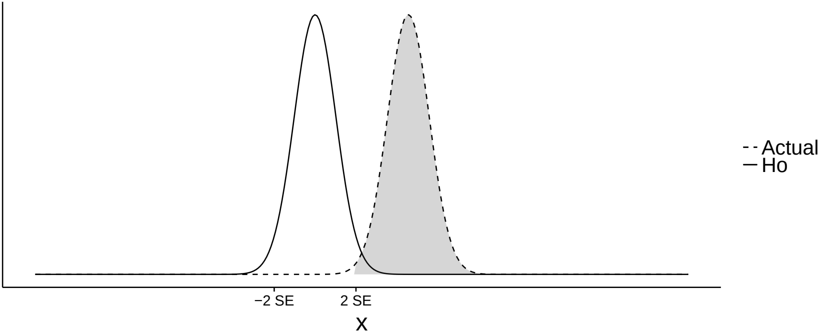


Given these sampling distributions, what is the probability that we correctly reject the null hypothesis? This is statistical power.

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!


−2 SE
2 SE
x
Actual
Ho