CM 224 : Computational Genetics
Homework #2

Ronak Sumbaly
UID - 604591897

Problem 1 - Association Study at a Single SNP

Part A - Non-Centrality Parameter

$$\texttt{Non-Centrality Parameter} = \lambda_A \sqrt{N} = \frac{p_A^+ - p_A^-}{\sqrt{2p_{mean}(1-p_{mean})}}\sqrt{N} \tag{1}$$

Given p_A and relative risk (γ) we can compute true case frequency (p_A^+), true control frequency (p_A^-) and p_{mean} as follows,

$$p_A^+ = \frac{\gamma p_A}{(\gamma - 1)p_A + 1} \qquad p_A^- = p_A \qquad p_{mean} = \frac{p_A^+ + p_A^-}{2} \tag{2}$$

R Code for calculating Non-Centrality Parameter

```
non_centrality <- function(gamma, pa, N){
  p_plus <- (gamma * pa)/((gamma - 1) * pa + 1)
  p_minus <- pa
  p_mean <- (p_plus + p_minus) / 2
  lambda <- (p_plus - p_minus) / (sqrt(2 * p_mean *(1 - p_mean)))
  non_centrality_parameter = lambda * sqrt(N)
  return(non_centrality_parameter)
}

non_centrality(1.5, 0.05, 500)

## [1] 1.52396
```

Non-Centrality Parameter Results

500 individuals

γ / p_A	0.05	0.2	0.4
1.5	1.52396	2.70666	3.178209
2.0	2.756259	4.767313	5.423261
30	4.697591	7.784989	8.451543

1000 individuals

γ / p_A	0.05	0.2	0.4
1.5	2.155205	3.827795	4.494666
2.0	3.897938	6.741999	7.66965
3.0	6.643397	11.00964	11.95229

Part B - Power

$$\texttt{Power Equation} = \Phi(\Phi^{-1}(\alpha/2) + \lambda_A\sqrt{N}) + 1 - \Phi(-\Phi^{-1}(\alpha/2) + \lambda_A\sqrt{N}) \qquad \alpha = 0.05 \tag{3}$$

R Code for calculating Power

```
power <- function(gamma, pa, N, alpha=0.05){
  return(pnorm(qnorm(alpha/2) + non_centrality(gamma, pa, N)) + 1 -
         pnorm((-qnorm(alpha/2) + non_centrality(gamma, pa , N))))
}

power(1.5, 0.05, 500)

## [1] 0.331664
```

Power Results

500 individuals

γ / p_A	0.05	0.2	0.4
1.5	0.331664	0.7723779	0.8884346
2	0.7870708	0.9975024	0.9997332
3	0.9969058	1	1

1000 individuals

γ / p_A	0.05	0.2	0.4
1.5	0.5774172	0.9691072	0.9943728
2	0.9736868	0.9999991	1
3	0.9999986	1	1

Part C - Number of Individuals

Loop through various N values and find power which is $\sim .80$

```
number_individuals <- function(gamma, pa){
  individuals <- seq(1:2000)
  ind <- numeric()
  close_to_80 <- numeric()
  for (i in individuals)  {
    if (abs(power(gamma,pa,i) - .80) <= .005) {
      close_to_80 <- c(close_to_80, abs(power(gamma,pa,i) - .80))
      ind <- c(ind, i)
    }
  }
  return (ind[which.min(close_to_80)])
}

number_individuals(1.5, 0.05)

## [1] 1690
```

Number of Individuals Results

γ / p_A	0.05	0.2	0.4
1.5	1690	536	389
2	517	173	133
3	178	65	55

Problem 2 - Unbalanced Cases and Control

Part A

Given,
In a balanced study, $\texttt{NCP} = \lambda_A\sqrt{N}$

In an unbalanced study, $\texttt{NCP} = \lambda_A\sqrt{\frac{2(N^+N^-)}{N^+ + N^-}}$

Assuming three times the number of cases as controls. Rewriting the above equations

$$N^+ = 3N^-$$
$$N^- = (1/2)N'$$
$$N^+ = (3/2)N'$$

Replacing $N^+ \texttt{and} N^-$ in the NCP of unbalanced study and set it equal to NCP of balanced study,

$$\lambda_A\sqrt{N} = \lambda_A\sqrt{\frac{2(N^+N^-)}{N^+ + N^-}} = \lambda_A\sqrt{\frac{2((3/2)N'(1/2)N')}{2N'}}$$
$$\sqrt{N} = \sqrt{\frac{3N'}{4}}$$
$$N' = \frac{4N}{3}$$

Hence, the study must be of size $\frac{4N}{3}$ to achieve same power as balanced study with N individuals.

Part B

Given,
$\texttt{Number of cases} = N^+/2$
$\texttt{Number of controls} = \infty$
Similar to Part A, equate NCP of both balanced and unbalanced studies and substitute above values,

$$\lambda_A\sqrt{N} = \lambda_A\sqrt{\frac{2(N^+N^-)}{N^+ + N^-}}$$
$$\lambda_A\sqrt{N} = \lambda_A\sqrt{2N^+ \lim_{N^- \to \infty} \frac{N^-}{(2N^+) + N^-}}$$
$$\lim_{N^- \to \infty} \frac{N^-}{(2N^+) + N^-} = \lim_{N^- \to \infty} \frac{N^-}{N^-(2N^+/N^- + 1)}$$
$$= 1 \qquad \texttt{law of infinity}$$
$$\sqrt{N} = \sqrt{2N^+}$$
$$N = 2N^+$$

Part C

We know that both the cases and control follow the distribution,

$$\hat{p^+} \sim N\left(p_A^+, \frac{p_A^+(1-p_A^+)}{N^+}\right)$$
$$\hat{p^-} \sim N\left(p_A^-, \frac{p_A^-(1-p_A^-)}{N^-}\right)$$

Taking the difference between the observed frequency between cases and controls,

$$\hat{p^+} - \hat{p^-} \sim N\left(p_A^+ - p_A^-, \frac{p_A^+(1-p_A^+)}{N^+} + \frac{p_A^-(1-p_A^-)}{N^-}\right)$$
$$\hat{p^+} - \hat{p^-} \sim N\left(p_A^+ - p_A^-, \frac{N^- p_A^+(1-p_A^+) + N^+ p_A^-(1-p_A^-)}{N^+N^-}\right)$$

Following the approximation,

$$N^- p_A^+(1-p_A^+) + N^+ p_A^-(1-p_A^-) \approx (N^- + N^+)(p_A(1-p_A))$$

Substituting we get,

$$\hat{p^+} - \hat{p^-} \sim N\left(p_A^+ - p_A^-, \frac{(N^- + N^+)p_A(1-p_A)}{N^+N^-}\right) \qquad p_A = \frac{p_A^+ + p_A^-}{2}$$

Since we want the variance to be 1, divide the above equation by the standard deviation (square root of variation)(using rule of normal distribution),

$$S_A = \frac{\hat{p^+} - \hat{p^-}}{\sqrt{\frac{(N^- + N^+)\hat{p_A}(1-\hat{p_A})}{N^+N^-}}} \sim N\left(\frac{p_A^+ - p_A^-}{\sqrt{\frac{(N^- + N^+)p_A(1-p_A)}{N^+N^-}}}, 1\right)$$
$$S_A = N\left(\frac{p_A^+ - p_A^-}{\sqrt{\frac{2(N^- + N^+)p_A(1-p_A)}{2N^+N^-}}}, 1\right)$$
$$S_A = N\left(\frac{p_A^+ - p_A^-}{\sqrt{2p_A(1-p_A)}}\sqrt{\frac{2N^+N^-}{N^- + N^+}}, 1\right)$$
$$\lambda_A = \frac{p_A^+ - p_A^-}{\sqrt{2p_A(1-p_A)}}$$

$$\texttt{NCP} = S_A = \lambda_A\sqrt{\frac{2N^+N^-}{N^- + N^+}} \qquad (Answer)$$