STATS530 - HW 1

Mateusz Grobelny

Problem 1

Under a Case-control study the cases and controls proportions are chosen deliberately which may not reflect the frequencies of how cases actually occur in the population.

So: If N = population
and N = A+B+C+D
then in the actual population:

	Yi =1	**Yi = 0**	**Total**
Xi =1	A	B	A + B
Xi = 0	C	D	C + D
Total	A + C	C + D	N Popsize

This means that when taking a sample of the population:
Xi = Exposure
Yi = Disease state

	Yi =1	**Yi = 0**	**Total**
Xi =1	a	b	a + b
Xi = 0	c	d	c + d
Total	a + c or n1	c + d or n0	n (Sample of population)

In cross sectional sample:

a = A/N *n
b = B/N *n
c = C/N *n
d = D/N *n

In cross sectional the risk is directly related to the main population as it is calculated as:

```
risk = (# of type of individual) / (# whole pop) ∗ (# of ind
```

In case control:

a = n1/(A+C) *A
b = n0/(B+D) * B
c = n1/(A+C) * C
d = n0/(B+D) * D

In case control sampling the risk for a given group is determined from the total individuals present in the sampling which is not directly related to the risk in the actual population.

Furthermore, when substituting in the values for a,b,c,d with the equivalent cross sectional values the fractions simplify back to the original risk difference equations but when substituting in with case control samples the fractions do not simplify back to the original risk difference equations.

Problem 2

Yes, it is possible to estimate B1 since its possible to estimate odds ratio from case - control sampling:

```
B1 = log((A/B)/(C/D)) = log((P1/(1-P1))/(P2/(1-P2))
```

Problem 3

```
$ head /proc/cpuinfo
processor    : 0
vendor_id    : GenuineIntel
cpu family    : 6
model        : 45
model name    : Intel(R) Xeon(R) CPU E5-2620 0 @ 2.00GHz
stepping    : 7
microcode    : 0x710
cpu MHz        : 1508.984
cache size    : 15360 KB
physical id    : 0
```

```
$ head /proc/meminfo
MemTotal:       16221276 kB
MemFree:         7110612 kB
MemAvailable:   11063224 kB
Buffers:              32 kB
Cached:          3900360 kB
SwapCached:          128 kB
Active:          5694656 kB
Inactive:        2202820 kB
Active(anon):    3614788 kB
Inactive(anon):   658400 kB
```

Problem 4

```
$ plink --noweb --file hapmap1
```

Cases: 44

Controls: 45

genotyping rate: 0.99441

Problem 5

```
# Start importing allele data
AA = 10
Aa = 25
aa = 81
Total_ind = sum(AA,Aa,aa)

# Find total allels
total_A = AA*2 + Aa
total_a = aa*2 + Aa
total_allels = total_a +total_A

# Find p and q
p = total_A/total_allels
q = 1-p

print(p)
print(q)

# Find expected frequencies
expected_AA = p**2*Total_ind
expected_Aa = 2*p*q*Total_ind
expected_aa = q**2*Total_ind

expected_aa
expected_Aa
expected_AA

# Make a table of observed vs expected
```

```
observed = c(AA,Aa,aa)
expected = c(expected_AA,expected_Aa,expected_aa)
ob_ex_table = cbind(observed,expected)
ob_ex_table

##
# Where r is the number of populations, and c is the number
# DF = (r - 1) * (c - 1)
DF = (2 - 1) * (2 - 1)

# perform test
test_stat = 0
for (i in 1:3){
  test_val = (ob_ex_table[i,1] - ob_ex_table[i,2])**2 /(ob_e
  test_stat = test_stat + as.numeric(test_val)
}
test_stat

# Get p value
pchisq(test_stat,df=DF,lower.tail = FALSE)
```

Chi-square Result:

```
p = 0.0008171198
```