Spring 2022 Bio220 Lab Midterm 2

2022-05-13

- Use the `mt2answer.R` file for your answers.
- Exam consists of 3 questions. First 2 questions are straight-forward applications of course content. Last question requires a deeper understanding of the material.
- There is no data file for this exam.
- Exam duration is **2 hours**.

Question 1 (40 points)

In a study of migratory waterbirds' habitat loss/shift due to climate change, 36 individuals from a *Sandpiper* species is traced to locate their breeding sites. Each individual's breeding site is categorized according to its latitude. 4 categories (A to D) of latitude ranges are defined that span northern hemisphere, where A corresponds to lower and D to higher latitudes. Expected relative frequencies (from previous observations) for breeding sites in each latitude category is as follows:

A: 0.06, B: 0.11, C: 0.65, D: 0.18

New data from the 36 traced birds is given as:

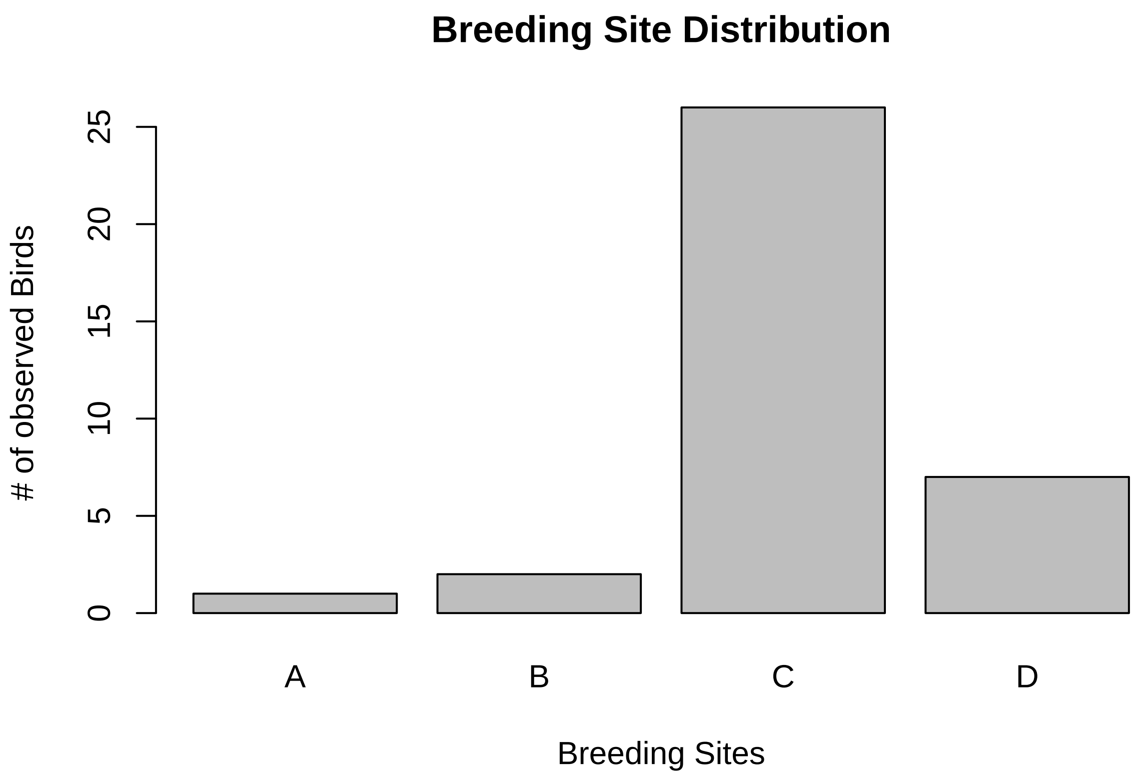
A: 1, B: 2, C: 26, D: 7

We want to test whether there is a latitudinal change in breeding sites of birds.

- Define given data in R as 2 vector objects.
- Using barplot, visualize the breeding site distribution of 36 traced birds. Write a title to the plot and label x and y axes. (Bonus: Use a named vector, so that name of the corresponding category is written under each bar.)
- Which test should be used? Check the assumption of the test and modify the data if necessary. (You may modify the data by manually defining new data objects instead of subsetting original objects.)
- State the null and alternative hypotheses.
- Perform the test and based on the output draw a conclusion about change of breeding sites.

```
# 1st part
q1_expected = c(A = 0.06, B = 0.11, C = 0.65, D = 0.18 )
q1_observed = c(A = 1, B = 2, C = 26, D = 7 )

# 2nd part
barplot(q1_observed, main = "Breeding Site Distribution",
        xlab = "Breeding Sites", ylab = "# of observed Birds")
```



```
# 3rd part
# chi-squared goodness-of-fit test should be used
q1_test = chisq.test(q1_observed, p = q1_expected)
```

```
## Warning in chisq.test(q1_observed, p = q1_expected): Chi-squared approximation
## may be incorrect
```

```
q1_test$expected
```

```
##     A     B     C     D
##  2.16  3.96 23.40  6.48
```

```
# 2 categories (A and B) have expected numbers below 5, we need to merge them
q1_exp_mod = c(q1_expected[1] + q1_expected[2], q1_expected[3], q1_expected[4])
q1_exp_mod
```

```
##    A    C    D
## 0.17 0.65 0.18
```

```
q1_obs_mod = c(q1_observed[1] + q1_observed[2], q1_observed[3], q1_observed[4])
q1_obs_mod
```

```
##  A  C  D
##  3 26  7
```

```
# 4th part
# H_0: data comes from expected population
# H_A: data is from a different population / doesn't fit expected population

# 5th part
chisq.test(q1_obs_mod, p = q1_exp_mod)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  q1_obs_mod
## X-squared = 1.9212, df = 2, p-value = 0.3827
```

```
# pval is greater than 0.05
# test is not statistically significant
# fail to reject H_0
# no evidence that data is from a different population
```

Question 2 (40 points)

As part of a genetic association study, effect of a *recessive risk allele* on a *disease phenotype* is studied. We want to know whether subjects having `aa` genotype have a **higher risk** of having the disease phenotype. Data from 28 subjects is as follows:

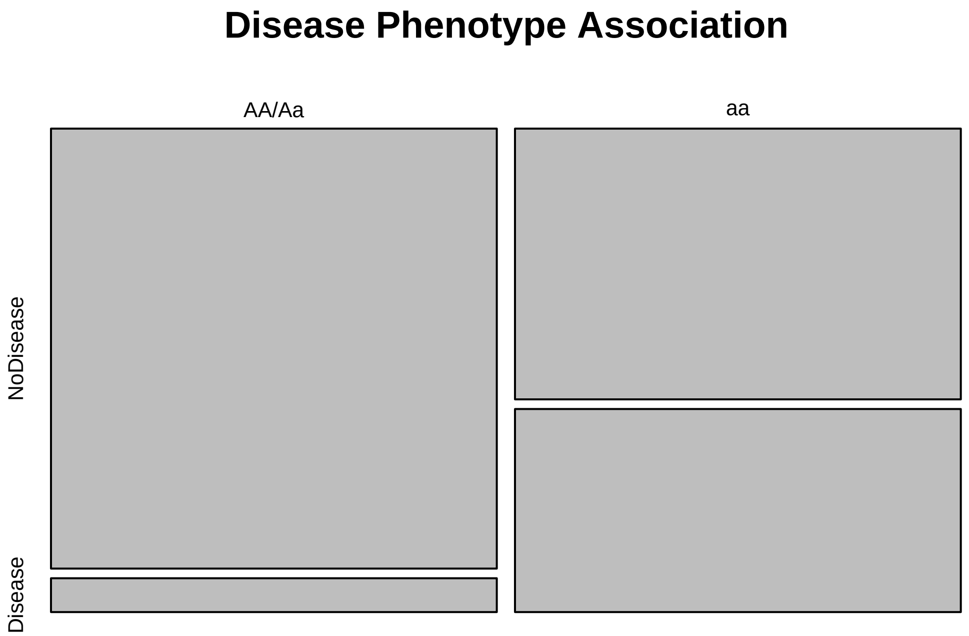
	NoDisease	Disease
AA/Aa	13	1
aa	8	6

- Define the data in R as a data frame. Make sure that the object has column names and row names.
- Visualize the data with a mosaic plot.
- Decide on the statistical test to use.
- State null and alternative hypotheses. Be careful about the alternative hypothesis.
- Perform the test and based on the output draw a conclusion about whether this allele may be associated with higher disease risk or not.

```
# 1st part
q2_data = data.frame(NoDisease = c(13, 8), Disease = c(1,6))
rownames(q2_data) = c("AA/Aa", "aa")
q2_data
```

```
##       NoDisease Disease
## AA/Aa        13       1
## aa            8       6
```

```
# 2nd part
mosaicplot(q2_data, main = "Disease Phenotype Association")
```



```r
# 3rd part
# we are asked to do a 1-sided test, so Fisher's test should be used

# 4th part
# aa genotype having higher disease risk means AA/Aa genotypes having
# lower disease risk,
# aa-disease and AA/Aa-NoDisease categories should have higher numbers
# this asked question should be the alternative hypothesis:
# H_0: no association btw genotype and disease, OR = 1
# H_A: AA/Aa group should have higher number of NoDisease, OR > 1

# 5th part
fisher.test(q2_data, alternative = "greater")
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  q2_data
## p-value = 0.03841
## alternative hypothesis: true odds ratio is greater than 1
## 95 percent confidence interval:
##  1.110927      Inf
## sample estimates:
## odds ratio
##   8.995715
```

```
# pval is less than 0.05
# test is statistically significant
# reject H_0
# we have evidence that aa genotype has higher disease risk
```

Question 3 (20 points)

We want to test germination rate of seeds from a seed bank by testing a sample of `30 seeds`. `0.8` is considered a good rate of germination, so we will use it as the `null hypothesis`. Alternative hypothesis is: `germination rate < 0.8`. Before performing it, we want to check power of the test. More specifically, we are interested in the probability of the test detecting it, if the population germination rate is 0.75. Rates down to 0.75 is acceptable to us, but we really want to be able to detect it, if it is 0.75 or lower.

There is a simulation code on the answer file for calculating power of the test for this case. Read and understand what it does and fill in the blanks within the code.

Logic of the simulation is as follows:

We want to know, how often we can detect that germination rate is lower than 0.8, when the true population germination rate is 0.75. We will draw many (=1000) random samples (of n=30) from a population with germination rate 0.75 and test each of them with a binomial test (with $H_0 : rate = 0.8, H_A : rate < 0.8$). Then count how many of these tests are significant.

```
# 1st part
x = 0.75
y = 0.8
n = 30
t = 1000

pval_dist = c()
for (i in 1:t){
  a_sample = sample(c("germ","nogerm"), size = n, replace = T, prob = c(x, 1-x))
  sample_germ_count = sum(a_sample=="germ")
  pval_dist[i] = binom.test(sample_germ_count, n = n, p = y, alternative = "less")$p.val
}
mean(pval_dist < 0.05)
```

```
## [1] 0.118
```

```
# 2nd: what is the last calculated value? power or type II error rate?
# power

# 3rd: is lower or higher values better?
# higher

# 4th: what kind of changes in x,y,n or t could improve this value?
# increase in n and increase in |x-y| would improve power. t would only change
# noise in our simulation.
```