Lab Assignment 02

Beatriz Gascón

Oct 25, 2018

Preamble

This assignment covers material from the following tutorials:

- Tutorial_07: Estimating proportions
- Tutorial_08: Binomial distribution
- Tutorial_09: Goodness of fit tests
- Tutorial_11: Contingency analysis
- Tutorial_12: Normal distribution

The assignment will be graded out of **12** (worth of each question is noted below). It is worth 3% of your final grade, and contributes to the "Lab" portion of your BIOL202 grade.

It is due **before** the beginning of your lab section during the week of October 22nd, 2018.

Late assignments will receive a **zero**, so do not attempt to submit your assignment immediately before your lab section.

Required packages

- `tigerstats`
- `binom`

Be sure to load these packages!

Required data

- "hostplant.csv" for Question 2
- "perfume.csv" for Question 3
- "plumage.csv" for Question 4

Load required datasets:

```
hostplant <- read.csv(url("https://people.ok.ubc.ca/jpither/datasets/hostplant.csv"), header = TRUE)
perfume <- read.csv(url("https://people.ok.ubc.ca/jpither/datasets/perfume.csv"), header = TRUE)
plumage <- read.csv(url("https://people.ok.ubc.ca/jpither/datasets/plumage.csv"), header = TRUE)
```

Reminder

Consult Tutorial_00 for instructions on how to prepare and submit you assignment.

Question 1 (1 mark)

No dataset is required for this question.

Among 38 students randomly sampled from a school, 9 had red hair. What is your best estimate of the proportion of red-haired students at the school? Provide an appropriate 95% confidence interval with your estimate.

```
pop.size <- 38

number.red <- 9

true.prop.red <- number.red/pop.size

true.prop.red
```

```
## [1] 0.2368421
```

The best estimate of the proportion of red-haired students at the school is of 0.24

```
ac.conf <- binom.confint(x = number.red,
                         n = pop.size,
                         conf.level = 0.95,
                         methods = "ac")
ac.conf
```

```
##          method x  n      mean   lower     upper
## 1 agresti-coull 9 38 0.2368421 0.12793 0.3940752
```

According to the calculationg above, the 95% confidence interval for the estimate is 0.13 ≤ p ≤ 0.39, which encompasses the true proportion.

Question 2 (2 marks)

Use the "hostplant.csv" dataset for this question.

A researcher suspected that planthoppers (a type of insect) detect their preferred host plant species by sensing volatile compounds produced by the plant. She conducted a well-designed experiment involving 23 independent trials in which the insect chose either the preferred host plant species (category `preferred` in the variable `hostchoice`) or a different host plant species (category `different` in the same variable `hostchoice`). Does this evidence support her suspicions?

Conduct an appropriate hypothesis test, and include the following in your answer:

- Appropriately worded null and alternative hypothesis statements
- Your chosen level of α
- an appropriate statistical test, and justification for your choice
- a concluding statement that includes your observed test statistic, an associated *P*-value, and an appropriate 95% confidence interval

- you do **NOT** need to provide a graph of any sort

```
inspect(hostplant)
```

```
##
## categorical variables:
##         name  class levels  n missing
## 1 hostchoice factor      2 23       0
##                                    distribution
## 1 preferred (78.3%), different (21.7%)
```

- Null and alternative hypothesis:

H0: The proportion of planthoppers choosing the "preferred" plant = 0.5

HA: The proportion of planthoppers choosing the "preferred" plant is not 0.5

- Level of alpha:

α = 0.05

- Statistical test:

The response variable is nominal (categorical) and it doesn't represent an association between two variables. Since it has two categories/ levels, the most appropriate test to apply is the Binomial test.

```
sum(hostplant$hostchoice == "preferred")
```

```
## [1] 18
```

```r
addmargins(xtabs(~ hostchoice, hostplant))
```

```
## hostchoice
## different preferred       Sum
##         5        18        23
```

```r
binom.result <- binom.test(~ hostchoice,
               data = hostplant,
               p = 0.5,
               success = "preferred",
               alternative = "two.sided",
               ci.method = "Agresti-Coull")
binom.result
```

```
##
##  Exact binomial test (with Agresti-Coull CI)
##
## data:  hostplant$hostchoice  [with success = preferred]
## number of successes = 18, number of trials = 23, p-value = 0.01062
## alternative hypothesis: true probability of success is not equal to
0.5
## 95 percent confidence interval:
##  0.5766743 0.9076511
## sample estimates:
## probability of success
##              0.7826087
```

- Conclusion:

Plant-hoppers chose the "preferred" plant at a significantly higher propotion than the "different"plant (n = 23; observed proportion of choice of "preferred" plant = 0.78; Binomial test; P-value = 0.01; Agrest-coull 95% confidence interval: $0.577 \leq p \leq 0.908$)

Question 3 (3 marks)

In a follow-up study to the one described in Question 2, the researcher suspected that perfume affected the ability of the planthoppers to detect their preferred host plant species. She conducted a well-designed experiment in which 23 independent planthoppers were randomly assigned to a "control" group, and 23 were randomly assigned to a "perfume" group. The control group underwent the trials as described in Question 2, whereas the perfume group had perfume introduced to the experimental apparatus.

The "perfume" dataset includes a variable `hostchoice` (similar to the preceding question) and a variable `treatment`, which indicates which treatment was applied.

Conduct an appropriate hypothesis test, and include the following in your answer:

- Appropriately worded null and alternative hypothesis statements
- Your chosen level of α
- an appropriate graph including figure caption, and one or two lines describing what you see
- an appropriate statistical test, and justification for your choice
- a concluding statement that includes your observed test statistic, an associated *P*-value, and an appropriate 95% confidence interval

```
inspect(perfume)
```

```
##
## categorical variables:
##         name  class levels  n missing
## 1 hostchoice factor      2 46       0
## 2  treatment factor      2 46       0
##                                   distribution
## 1 preferred (65.2%), different (34.8%)
## 2 control (50%), perfume (50%)
```

- Null and alternative hypothesis:

H0: There is no association between the ability of the planthoppers to detect their preferred host plant species and perfume

HA: There is an association between the ability of the planthoppers to detect their preferred host plant species and perfume

- Level of alpha:

α = 0.05

- Statistical test :

The response variable is nominal(categorical) and it shows an association between two variables, where each one of them has exactly two categories and does not deal with health outcomes for people. The appropriate test to be conducted is, therefore, the Fisher's Exact test.

```
plant.hopperTable <- xtabs(~ treatment + hostchoice,
                data = perfume)

plant.hopperTable
```

```
##           hostchoice
## treatment different preferred
##   control         4        19
##   perfume        12        11
```

```
mosaicplot(t(plant.hopperTable),
      color = c("blue", "goldenrod1"),
```

```
      cex.axis = 0.9,
      xlab = "Treatment",
      ylab = "Condition",
      main = "")
```

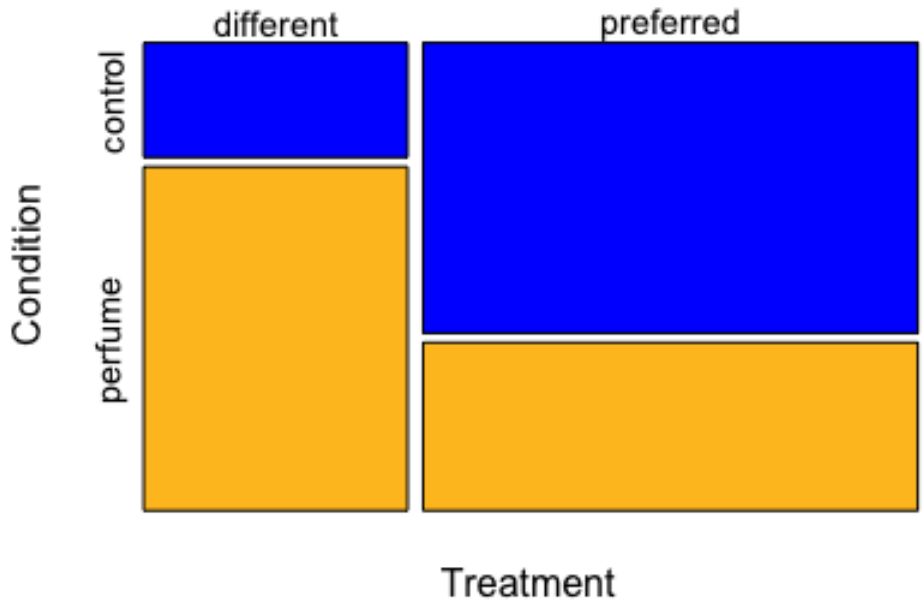

Figure 1: The relative frequency of hostchoice among study subjects (planthoppers) who were given a choice of their preferred plant or a different plant without perfume (n = 23) as well as their preferred plant or a different plant with perfume (n= 23).

```
planthoppers.fischer.results <-  fisher.test(plant.hopperTable)

planthoppers.fischer.results #Fischer's Exact test

##
##  Fisher's Exact Test for Count Data
##
## data:  plant.hopperTable
## p-value = 0.02868
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.03743948 0.87032431
## sample estimates:
```

```
## odds ratio
##  0.2005409
```

- Conclusion:

There is evidence that the probability of choosing a hostplant differs depending on wetheror not the plant has perfume (Fisher's Exact Test; P-value = 0.029; odds ratio = 0.20; 95% CI: 0.037 - 0.870).

Question 4 (4 marks)

Use the "plumage" dataset for this question.

A researcher suspected that plumage colour (i.e. the colour of the feathers) in a bird species she studied depended on the habitat in which the bird nested. The species exhibits three different colour morphs, blue, brown, and red. She randomly sampled 30 birds in each of 2 habitat types, forest and grassland, and recorded the plumage colour of each bird. Is there evidence to support her suspicions?

Conduct an appropriate hypothesis test, and include the following in your answer:

- Appropriately worded null and alternative hypothesis statements
- Your chosen level of α
- an appropriate graph including figure caption, and one or two lines describing what you see
- an appropriate statistical test, and justification for your choice
- a check of any assumptions of the test
- a concluding statement that includes your observed test statistic, an associated *P*-value
- you do **NOT** need to include a confidence interval in your concluding statement

```
inspect(plumage)
```

```
##
## categorical variables:
##      name  class levels  n missing
## 1  colour factor      3 60       0
## 2 habitat factor      2 60       0
##                                   distribution
## 1 blue (48.3%), brown (30%) ...
## 2 forest (50%), grassland (50%)
```

- Null and alternative hypothesis:

H0: There is no association between the habitat in which the bird nests and its plumage colour

HA: There is an association between the habitat in which the bird nests and its plumage colour

- Statistical test:

The response variable is nominal(categoreical) and it shows associations between two variables which do not have exactly 2 categories each, therefore the appropriate statistical test would be the X squared contingency test (test of indepence)

- Level of alpha:

α = 0.05

```
birdTable <- xtabs(~ habitat + colour, data = plumage )

birdTable # Contingency table
```

```
##            colour
## habitat     blue brown red
##   forest      17     8   5
##   grassland   12    10   8
```

```
birdTable.rel.freq <- prop.table(birdTable, margin = 2)

birdTable.rel.freq #corresponding relative frequencies
```

```
##            colour
## habitat          blue     brown       red
##   forest    0.5862069 0.4444444 0.3846154
##   grassland 0.4137931 0.5555556 0.6153846
```

```
mosaicplot(t(birdTable.rel.freq),
      col = c("blue", "goldenrod1"),
      cex.axis = 0.9,
      xlab = "Plumage colour",
      ylab = "Habitat",
      main = "") #Visualizing mosaic plot
```

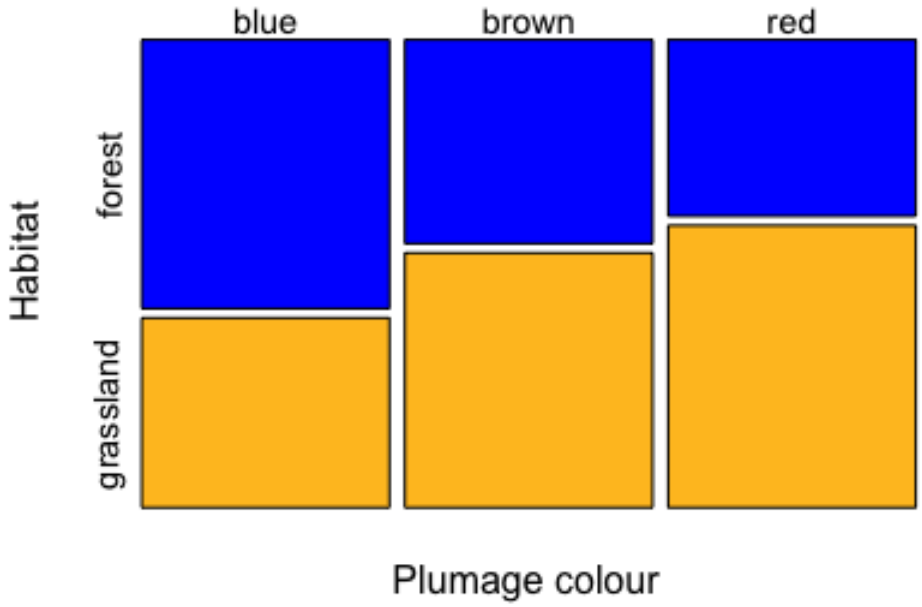

Figure 2: Relative frequency of plumage colour among 60 birds from different habitats

```
bird.chisq.results <- chisq.test(birdTable)

names(bird.chisq.results)

## [1] "statistic" "parameter" "p.value"   "method"    "data.name" "obs
erved"
## [7] "expected"  "residuals" "stdres"

bird.chisq.results$expected

##            colour
## habitat     blue brown red
##   forest    14.5     9 6.5
##   grassland 14.5     9 6.5

bird.chisq.results

##
##  Pearson's Chi-squared test
##
```

```
## data:  birdTable
## X-squared = 1.7766, df = 2, p-value = 0.4114
```

- Conclusion:

The probability of being of a certain plumage colour is not significantly associated with the habitat (χ2 contingency test; df = 2; χ2 = 1.78; P > 0.411). Based on Figure 2, the probability of having a certain plumage colour does not change substantially depending on the habitat.

Question 5 (2 marks)

No dataset is required for this question.

You do **NOT** need to provide a graph of any sort for these questions.

Hatchling weight in the general population of chickadees (a type of bird) is normally distributed with a mean of 8.2g and a standard deviation of 1.10g.

(a) Calculate the probability of a randomly sampled hatchling having a weight of 7.3g or less.

```
pnormGC(7.3,
     region = "below",
     mean = 8.2,
     sd = 1.1,
     graph = FALSE)
```

```
## [1] 0.2066267
```

the probability of a randomly sampled hatchling having a weight of 7.3g or less is of 0.21, meaning Pr[weight < 7.3] = 0.21

(b) Calculate the probability of a randomly sampled hatchling having a weight between 8.4 and 9.0g.

```
prob_above_8.4 <- pnormGC(8.4,
     region = "above",
     mean = 8.2,
     sd = 1.1,
     graph = FALSE)

prob_below_9.0 <- pnormGC(9.0,
                region = "below",
                mean = 8.2,
                sd = 1.1,
                graph = FALSE)
```

```r
prob_between_8.4_and_9.0 <- prob_below_9.0 - prob_above_8.4

prob_between_8.4_and_9.0
```

```
## [1] 0.3386078
```

the probability of a randomly sampled hatchling having a weight between 8.4g and 9.09 is of 0.34, meaning Pr[8.0 < weight < 9.0] = 0.34