Midterm Key

Heath Blackmon

3/09/2019

For all questions give the shortest possible complete answers **one or to two sentences**

Section 1 (20 points)

What is a p-value?

A p-value is the probability that we would observe a result (a statistic) as extreme or more extreme as the one from our sample assuming our null hypothesis is true

What is the differnce in a statistic and a parameter?

A parameter is the value of a variable in a population. While a statistic is our estimate of that value based on a sample.

Describe one of the causes of the reproducibility crisis.

File drawer problem: people don't publish negative results.

p-hacking: Trying lots of tests and variables till something is significant, and then reporting the results as though that was the one test that was done.

small sample size: leads to more variation in results and low power.

pressure to publish: people have to publish so they engage in small amounts of some or all of the above in an effort to get high impact publications

statistical ignorance: people are poorly trained and make mistakes in their application of statistical tests.

maleficence: Some people suck, and they fabricate or manipulate data for their own personal gain rather than the advancement of science.

data dredging: Testing many possible explanatory variables and then reporting the significance of the ones that you find show an interesting pattern (significance).

Why do we have to correct for multiple comparisons?

Every test has a possibility of generating a false positive. To maintain an overall false positive rate of 5% (arbitrary but widely accepted) we must account for this and require a more significant result (extreme p-value) to consider the result significant.

In Bayesian analyses what is a prior used for?

A prior allows us to incorporate our biological knowledge into our analysis and update our beliefs.

Section 2 (20 points)

What is the default null hypothesis of a two sample t-test?

The two samples come from the same population. The mean of the populations being sampled is the same.

What is the null hypothesis of an ANOVA?

The means of the groups being tested is the same.

You perform an experiment where chickens are fed corn, mixed seed, and a commercial diet. The chickens have final mean weights of 0.53, 0.69, and 0.41 kilograms respectively. You run an ANOVA on this data and get a p-value of 0.0012. What is your interpretation of this result?

One of the feeds being tested leads to significantly different weights. We don't know if they all or even which ones differ.

Section 3 (20 points)

Report the probability of observing 18 or more retrogenes (offspring) on chromosome 1. Download the data (retrogene.csv) analyze it and describe what you did, the result, and your inference.

```
# read the data
dat <- read.csv("retrogenes.csv", as.is=T)

# what is the observation of interest
obs <- 18

# chromosome size will define the probabilities of observing a
# retrogene on a chromosome:
probs <- dat$Size

# now let set up a result vector to store our null distribution in
results <- c()

# lets get our sample size from the empirical data
samp.size <- sum(dat$number.of.offspring.genes)

# now lets simulate our null distribution with a loop
# I'm going to do a lot to have a precise and accurate
# estimate of the true p-value
iter <- 100000
for(i in 1:iter){
  if(i %% 10000 == 0) print(i)
  retros <- sample(dat$chrom.num,
                   size = samp.size,
                   prob = probs,
                   replace = T)
  retros <- sum(retros == "1")
  results[i] <- retros
}
```

```
## [1] 10000
## [1] 20000
## [1] 30000
## [1] 40000
## [1] 50000
## [1] 60000
## [1] 70000
## [1] 80000
## [1] 90000
## [1] 100000
```

```r
# now lets calculate a p-value
pval <- sum(results >= obs)/iter
```

I ran a Monte Carlo fro 100000 generation to generate a null distribution and calculate the probability of observing 18 or more retrogenes on chromosome 1. I calculated an empirical p-value of *pval*. This tells me that if my null hypothesis is true that I would expect this result or even more Retrogenes approximately $round(pval)\%$ of the time.

Section 4 (20 points)

A researcher was intersted in whether differences in fishing practices in Canada and the US have an impact on the size of fish present on the shores of lake superior. Download the data (fish.csv) analyze it (using the simplest appropriate test) and describe what you did, the result, and your inference.

```r
# read the data
dat <- read.csv("fish.csv", as.is = T)
# get rid of a row NA in the data - though code would
# work without this so it is really just doing good clean work.
dat <- dat[!is.na(dat$US), ]
result <- t.test(x = dat$US, y = dat$CAN)
pval <- result$p.value
```

I ran a simple two sample T-test. The p-value for this test was *pval* suggesting that their is no significant difference in the size of fish from the US and Canadian samples.

Section 5 (20 points)

Download the data file fins.csv and make a plot that does a good job of illustrating the difference in data from males and females. The plot should be publication quality.

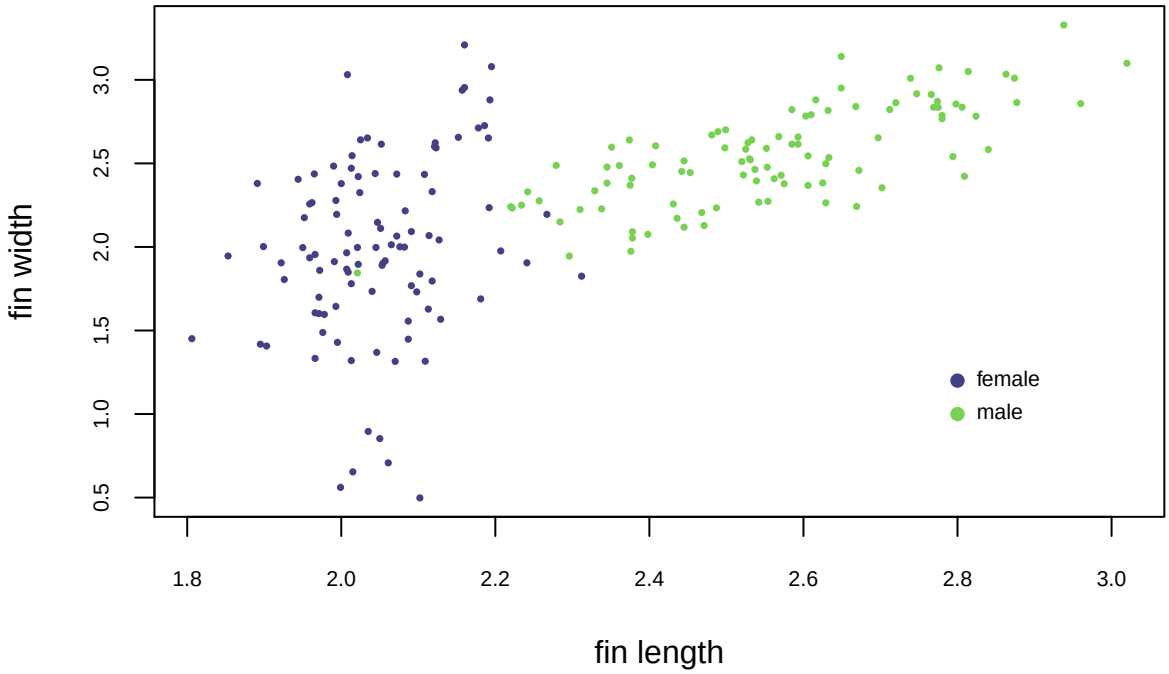
```r
# i'm going to use the viridis color package for colors
# because they are color blind friendly
library(viridis)
```

```
## Loading required package: viridisLite
```

```r
# read the data
dat <- read.csv("fins.csv")

# lets set up a vector of colors to represent sex
cols <- viridis(100)[c(20, 80)][dat$sex]
plot(x = dat$length,
     y = dat$width,
     col = cols,
     pch = 16,
     cex = .5,
     xlab = "fin length",
     ylab = "fin width",
     cex.axis = .7)
points(x=c(2.8,2.8),
       y=c(1.2,1),
       col=viridis(100)[c(20, 80)],
       pch=16)
```

```
text(x=c(2.8,2.8),
     y=c(1.2,1),
     c("female","male"),
     pos=4,
     cex=.7)
```



Extra Credit (20 points)

You develop a test based on expression profiles that can predict the developmental fate of two closely related neural progenitor cells as either motor or sensory (you know these are equally common in the pool of neural progenitor cells). The test has 99% probability of correctly predicting development of a motor neuron and a 50% probability of correctly predicting the development of a sensory neuron. What is the proability that a cell predicted by your test to develop into a motor neuron will actually develop as a motor neuron?

Lets remind ourselves of Bayes theorem first:

$p(a|b) = \frac{p(b|a)p(a)}{p(b)}$

```
# now lets plug in these values to variables

# this is the one we don't know
pab <-c()

# this is the probability of getting the result motor given it is a motor neuron

pba <- .99

# this is the probability of being a motor neuron

pa <- .5
```

```r
# this is the probability of getting the motor neuron result
# it has two components false positive and true positives

fp <- .5 * .5
tp <- .99 * .5

pb <- fp + tp

# now we can do the math

pab <- (pba*pa)/pb
```

There is a $round(pab * 100)\%$ chance that a cell that tests as positive for becoming motor neuron will become a motor neuron.