AM115 Exercise 1

Ali Crump

2/3/2020

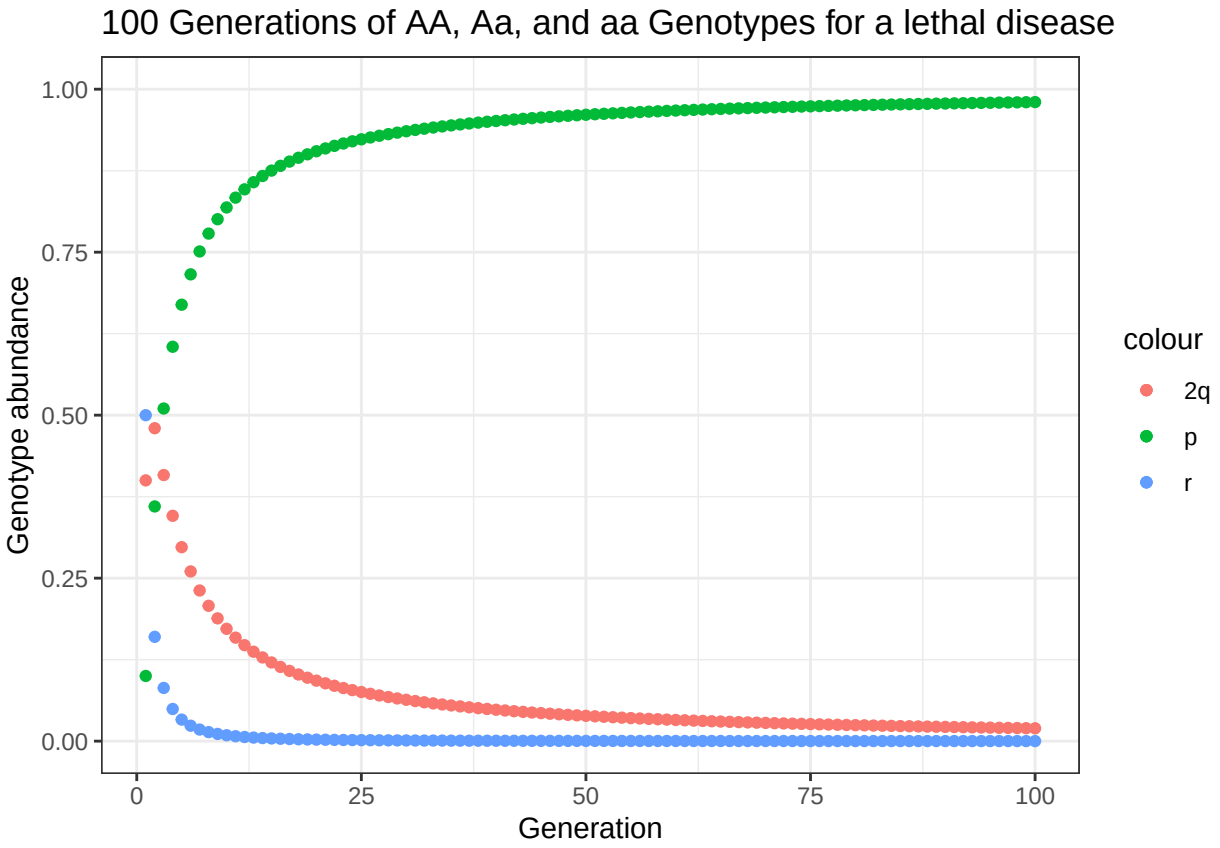
1. Hardy-Weinberg

b.

```
# Set initial conditions

p <- 0.1
q <- 0.2
r <- 0.5

# Define vectors

dominant <- 1:100
carrier <- 1:100
recessive <- 1:100

# Make initial condition first value of vector

dominant[1] <- p
carrier[1] <- 2*q
recessive[1] <- r

# Iterate through 100 generations of calculating p,q,r

for(i in 1:99 ){
  dominant[i+1] <- (p^2 + 2*p*q + q^2) / ((p+2*q)^2)
  carrier[i+1] <- (2*q*(p+q)) / ((p+2*q)^2)
  recessive[i+1] <- (q^2) / ((p+2*q)^2)
  p <- dominant[i+1]
  q <- carrier[i+1] / 2
  r <- recessive[i+1]
}

# join all together in one matrix

a <- cbind(dominant,carrier,recessive)
```

```
# Plot dominant, carrier, and recessive over 100 generations

a %>%
  as.data.frame(a) %>%
  ggplot(aes(x = 1:100)) +
  geom_point(aes(y = dominant, color = "p")) +
  geom_point(aes(y = carrier, color = "2q")) +
  geom_point(aes(y = recessive, color = "r")) +
  labs(x = "Generation", y = "Genotype abundance",
       title = "100 Generations of AA, Aa, and aa Genotypes for a lethal disease") +
```

```
ylim(0:1) +
theme(panel.grid.major = element_blank()) +
theme_bw()
```



2. Benford's Law

I'm from Pennsylvania so I decided to look at the data set containing Pennsylvania county populations over the past 10 years. I got the data from census.gov. The data appears to generally follow Benford's law, with 1 occurring about 25% of the time and 9 about 2% of the time. I was a bit suprised to see that 4 appears more often than 1 (26%). This could be explained by the fact that Allegheny County and Philadelphia County are the only counties with more than 1 million residents. Other counties in Pennsylvania are much smaller, as there aren't really any other major cities in Pennsylvania, and counties less than 20,000 are so small that they're very rare. This might also be explained by the fact that this data is for the same 67 counties over 10 years or so. We won't expect county populations to change very much from year to year so it's a relatively small sample with basically only 67 counties.

```
papop <- melt_csv("PEP_2018_PEPANNRES_with_ann.csv") %>%
  filter(data_type == "integer")
papop$row_num <- seq.int(nrow(papop))
papop <- papop %>%
  select(value, row_num) %>%
  mutate(value = as.numeric(value))
```

```
# Extract the first digit of the population
x <- extract.digits(papop$value, number.of.digits = 1) %>%
  # Count the number of times each digit appears
  group_by(data.digits) %>%
  count() %>%
  ungroup() %>%
  # Calculate the percentage of times each number appears
  mutate(percent = n / sum(n))
```

```
x %>%
  select(data.digits,percent) %>%
  mutate(percent = round(percent, digits=2)) %>%
  gt() %>%
  tab_header(title = "Benford's Law",
             subtitle = "PA County Populations") %>%
  cols_label(
    data.digits = "First Digit",
    percent = "Percent"
  )
```

Benford's Law
PA County Populations

First Digit	Percent
1	0.25
2	0.10
3	0.11
4	0.26
5	0.08
6	0.08
7	0.03
8	0.06
9	0.02

```r
x %>%
  # Make bar graph with digit on x and percent it appears on y
  ggplot(aes(x = data.digits, y = percent)) +
  geom_col() +
  # Add labels
  labs(x = "First Digit", y = "Percent", title = "Pennsylvania County Populations", subtitle = "Benford
  # Add x axis ticks
  scale_x_continuous(breaks = c(1:9)) +
  # Get rid of gray background
  theme_bw() +
  # Get rid of gridlines
  theme(panel.grid.major = element_blank())
```