HW1_wgeither

Warren Geither

8/29/2020

Problem 1

Primers Done

Problem 2 & 3

Saved to Github

a) I got my undergrad in mathematics with a minor in stats, so I had a fair amount of exopsure and use in R. For the past 2 years I worked as a Data Analyst at an Ad-tech company where I primarily used Python, SQL, and C#. That being said, in this class I would love to:

- Refamiliarize myself with linear regression & ANOVA in R
- Learn about parallel computing and how to connect to the university's supercomputer
- Learn about Monte Carlo procedures and Power as I'm not sure if this was covered in my undergrad curriculum

b)

$$\text{Binomial: } P(X = x|n, p) = \binom{n}{x} \cdot p^x \cdot (1-p)^{n-x}; x = 0, 1, 2, ..., n;\ 0 \leq p \leq 1 \tag{1}$$

$$\text{Discrete Uniform: } P(X = x|N) = \frac{1}{N}; x = 0, 1, 2, ..., N;\ N \in \mathbb{N} \tag{2}$$

$$\text{Exponential: } f(x|\beta) = \frac{1}{\beta} e^{-x/\beta},\ 0 \leq x \leq \infty,\ \beta > 0 \tag{3}$$

Performing Reproducible Research

1. For Every Result, Keep Track of How It Was Produced
2. Avoid Manual Data Manipulation Steps
3. Archive the Exact Versions of All External Programs Used
4. Version Control All Custom Scripts
5. Record All Intermediate Results, When Possible in Standardized Formats
6. For Analyses That Include Randomness, Note Underlying Random Seeds
7. Always Store Raw Data behind Plots
8. Generate Hierarchical Analysis Output, Allowing Layers of Increasing Detail to Be Inspected
9. Connect Textual Statements to Underlying Results
10. Provide Public Access to Scripts, Runs, and Results

Problem 4

```
# look at available R datasets
# library(help="datasets")

# assign stock data set Orange to data variable
data = Orange

# check out the data to see what we can plot
print(data)
```

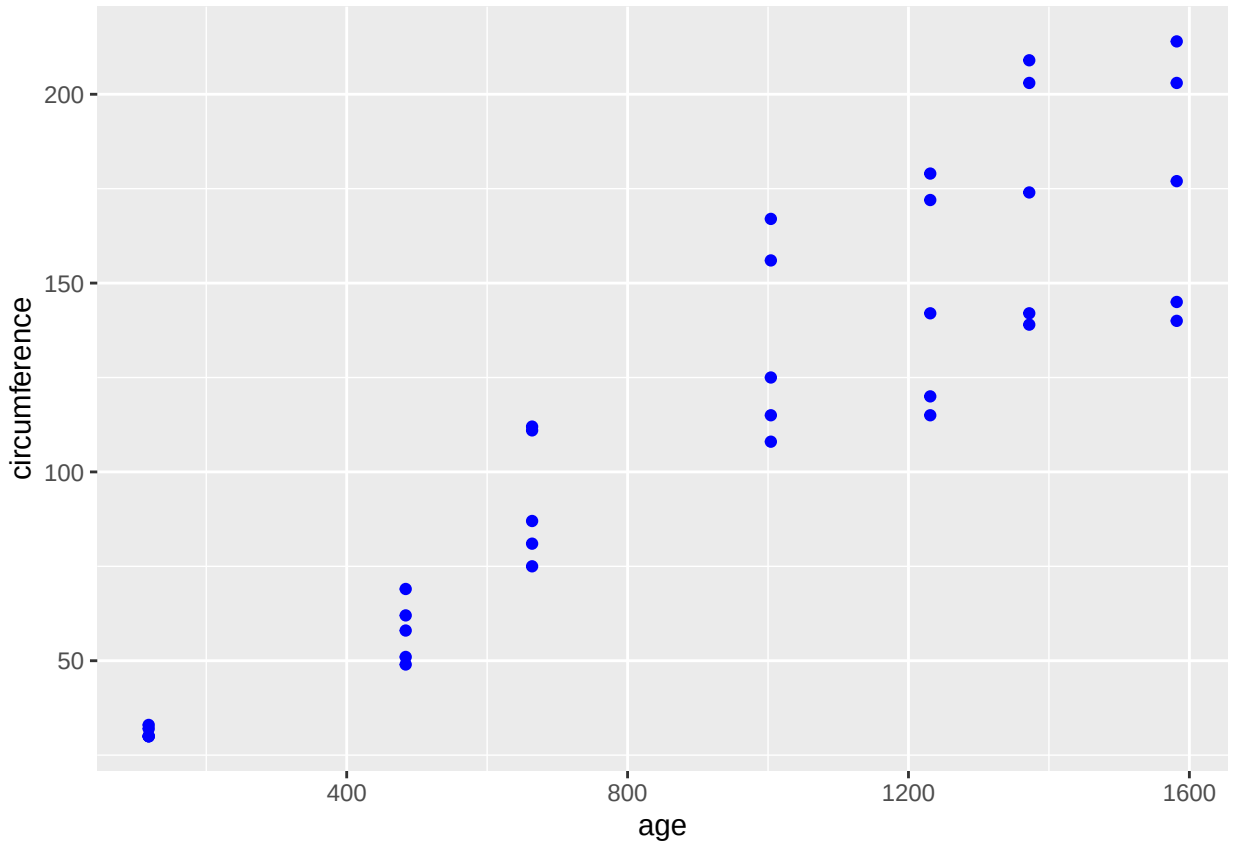
```
##    Tree  age circumference
## 1     1  118            30
## 2     1  484            58
## 3     1  664            87
## 4     1 1004           115
## 5     1 1231           120
## 6     1 1372           142
## 7     1 1582           145
## 8     2  118            33
## 9     2  484            69
## 10    2  664           111
## 11    2 1004           156
## 12    2 1231           172
## 13    2 1372           203
## 14    2 1582           203
## 15    3  118            30
## 16    3  484            51
## 17    3  664            75
## 18    3 1004           108
## 19    3 1231           115
## 20    3 1372           139
## 21    3 1582           140
## 22    4  118            32
## 23    4  484            62
## 24    4  664           112
## 25    4 1004           167
## 26    4 1231           179
## 27    4 1372           209
## 28    4 1582           214
## 29    5  118            30
## 30    5  484            49
## 31    5  664            81
## 32    5 1004           125
## 33    5 1231           142
## 34    5 1372           174
## 35    5 1582           177
```

```
# Read about the dataset
# ?Orange
```

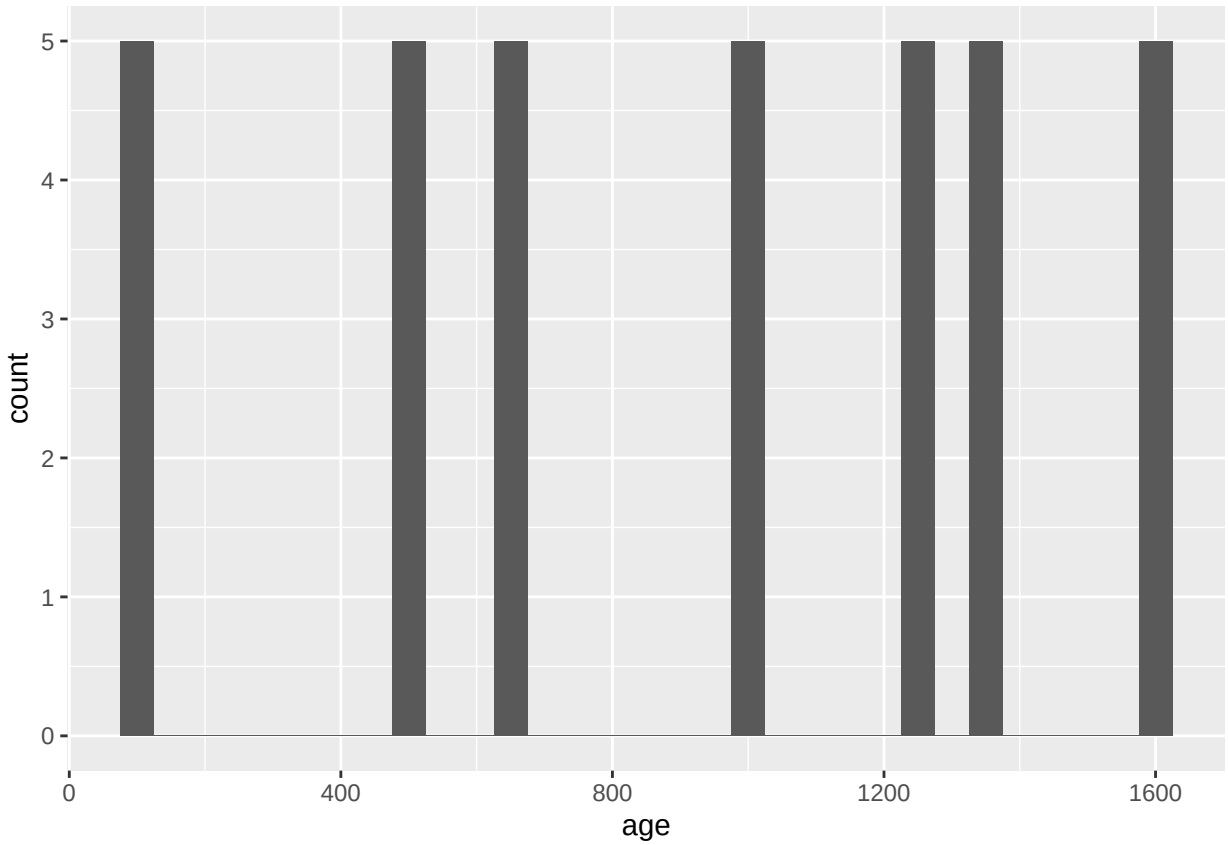
```
# install and load package ggplot2 if you did not already
# install.packages("ggplot2")
```

```
# initialize library
library("ggplot2")

# use ggplot package to create a scatterplot of age vs. circumference
ggplot(data) +
        geom_point(mapping = aes(x=age, y=circumference), color="blue")
```




```
# use ggplot package to create a histogram of ages of trees (days since 1968/12/31)
ggplot(data) +
        geom_histogram(mapping = aes(x=age), binwidth=50)
```


count
5
4
3
2
1
0
0
400
800
1200
1600
age