Chapter 4 Probability Distributions

Angelo LaCommare-Soto

2025-02-19

Section 1 - Importing Data

```
# set working directory for all chunks in this file (default working directory is wherever Rmd file is)
getwd()
```

```
## [1] "C:/Users/Angelo L/Documents/GitHub/BIOL710/RCode710/RCode/working_directory"
```

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```
# Importing data
cayo <- read.csv("cayo.csv",header=TRUE)

# viewing the structure of the data
str(cayo)
```

```
## 'data.frame':    1480 obs. of  8 variables:
##  $ dob            : chr  "1985-02-20T00:00:00Z" "1985-03-26T00:00:00Z" "1985-03-06T00:00:00Z" "1985-0
##  $ season         : int  1985 1985 1985 1985 1986 1986 1986 1987 1987 1987 ...
##  $ sex            : chr  "u" "u" "m" "m" ...
##  $ dod            : chr  "1985-02-20T00:00:00Z" "1985-03-26T00:00:00Z" "1985-05-28T00:00:00Z" "1986-0
##  $ mom_dob        : chr  "1981-06-13T00:00:00Z" "1981-03-05T00:00:00Z" "1981-01-01T00:00:00Z" "1981-0
##  $ age_at_delivery: num  3.69 4.06 4.18 4.04 4.75 ...
##  $ age_at_death   : num  0 0 0.2274 0.9178 0.0192 ...
##  $ treatment      : chr  "control" "control" "control" "control" ...
```

```
head(cayo)
```

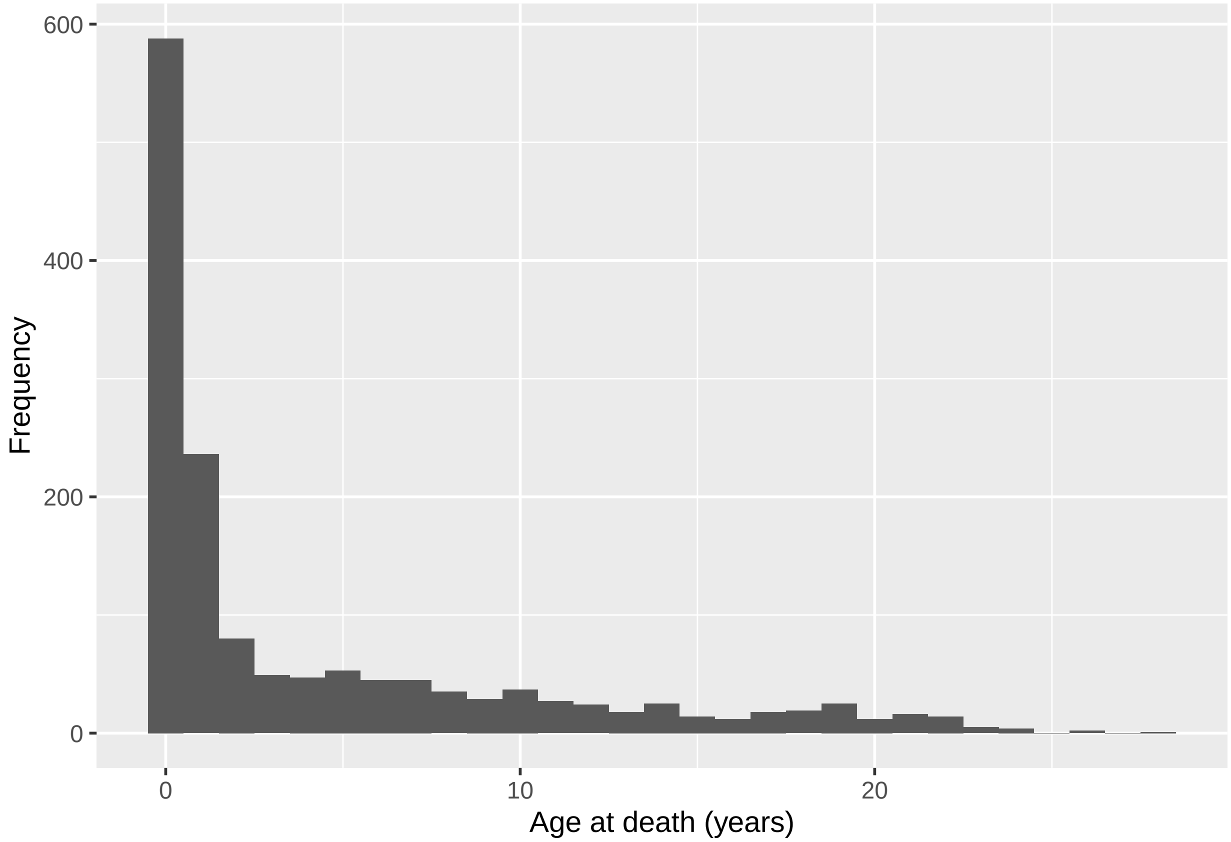
```
##                      dob season sex                  dod               mom_dob
## 1 1985-02-20T00:00:00Z   1985   u 1985-02-20T00:00:00Z 1981-06-13T00:00:00Z
## 2 1985-03-26T00:00:00Z   1985   u 1985-03-26T00:00:00Z 1981-03-05T00:00:00Z
## 3 1985-03-06T00:00:00Z   1985   m 1985-05-28T00:00:00Z 1981-01-01T00:00:00Z
## 4 1985-03-03T00:00:00Z   1985   m 1986-02-01T00:00:00Z 1981-02-16T00:00:00Z
## 5 1986-01-20T00:00:00Z   1986   u 1986-01-27T00:00:00Z 1981-04-24T00:00:00Z
## 6 1986-05-12T00:00:00Z   1986   m 1986-06-23T00:00:00Z 1980-03-16T00:00:00Z
##   age_at_delivery age_at_death treatment
## 1        3.693151   0.00000000   control
## 2        4.060274   0.00000000   control
## 3        4.178082   0.22739726   control
## 4        4.043836   0.91780822   control
## 5        4.745205   0.01917808   control
## 6        6.158904   0.11506849   control
```

Question Answers

a. The ‘cayo’ dataset has 1480 observations of 8 variables.
b. Given the question ‘Does age at death vary across age in females?’, we are interested in the ‘sex’ and ‘age_at_death’ variables.

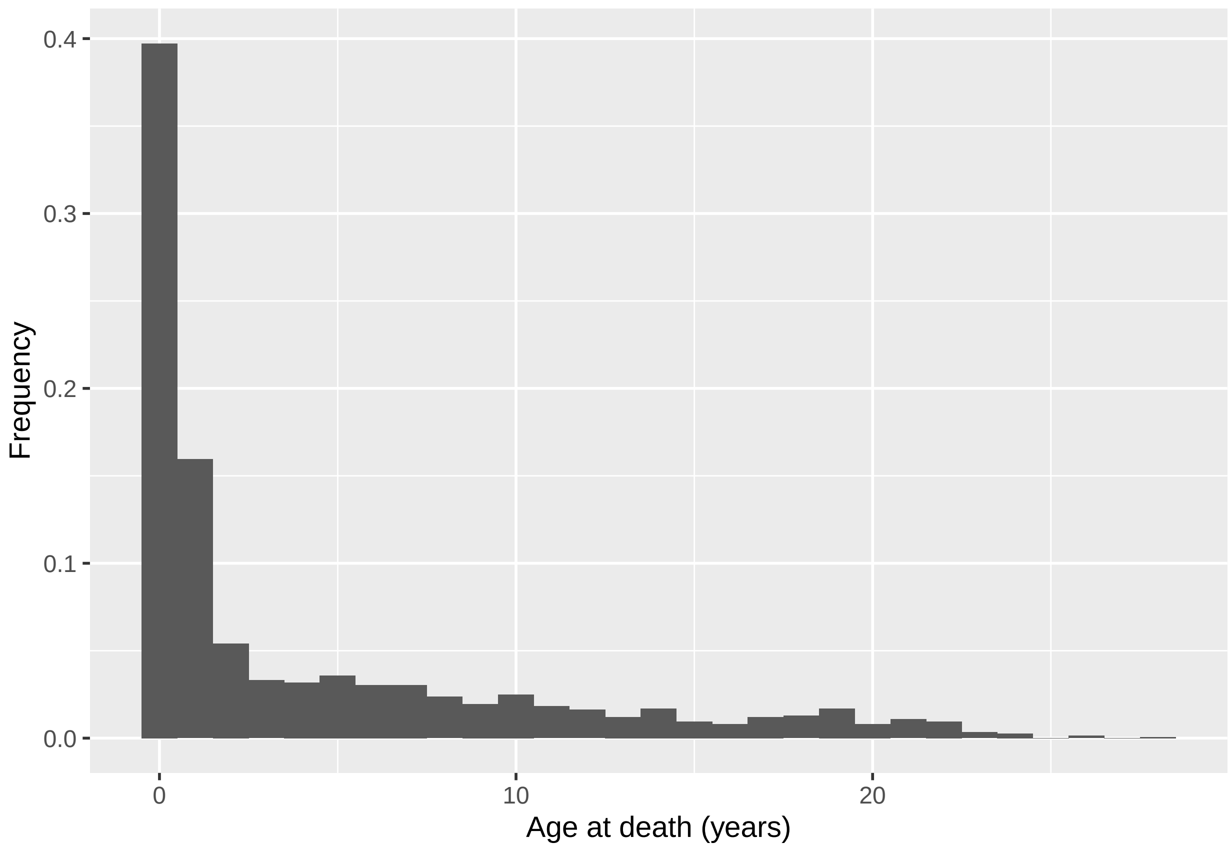
Section 2 - Infering Population Parameters from Random Samples

```r
# plotting counts, the function floor() rounds variable to the smallest whole number
p1 <- ggplot(cayo,aes(x=floor(age_at_death))) +
  geom_histogram(binwidth = 1) +
  xlab("Age at death (years)") +
  ylab("Frequency")
p1
```



```
# plotting proportions, ..count.. tallies the number of observations per unit (bin) on the x-axis from
p2 <- ggplot(cayo,aes(x=floor(age_at_death))) +
   geom_histogram(aes(y=..count../sum(..count..)),binwidth = 1) +
   xlab("Age at death (years)") +
   ylab("Frequency")
p2
```

```
## Warning: The dot-dot notation (‘..count..‘) was deprecated in ggplot2 3.4.0.
## i Please use ‘after_stat(count)‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```



```
# Estimating mean and sd population parameters for age at death
# mean age at death
mu <- mean(cayo$age_at_death)
mu
```

```
## [1] 4.684606
```

```
# standard deviation of age at death
sigma <- sd(cayo$age_at_death)
sigma
```

```
## [1] 6.178597
```

Question Answers

a. Age 0 (or newborn) monkeys experience a higher probability of deaths.
b. mu = 4.69 years of age is the population mean age of death for the monkeys. sigma = 6.18 years of age is the standard deviation, a measure of the precision of the mean age of death for the monkeys.

```
# checking the levels in the variable sex
levels(as.factor(cayo$sex))
```

```
## [1] "f" "m" "u"
```

Question Answers

a. The ‘sex’ categories are ‘f’ for female, ‘m’ for male, and ‘u’ for unknown.
b. Given that there are unknown categories of ‘sex’, we can only estimate a sampling distribution of sample means (average measure) and sample standard deviations (spread measure) for statistics involving the sexes.

```
# Estimating the mean age at death and its sd for females
# filtering by females
fem <- filter(cayo,sex=="f")
str(fem)
```

```
## 'data.frame':    737 obs. of  8 variables:
##  $ dob            : chr  "1987-03-02T00:00:00Z" "1987-02-21T00:00:00Z" "1987-12-24T00:00:00Z" "1988-0
##  $ season         : int  1987 1987 1988 1988 1987 1988 1985 1989 1989 1989 ...
##  $ sex            : chr  "f" "f" "f" "f" ...
##  $ dod            : chr  "1988-01-11T00:00:00Z" "1988-01-20T00:00:00Z" "1987-12-24T00:00:00Z" "1988-0
##  $ mom_dob        : chr  "1982-02-22T00:00:00Z" "1980-02-04T00:00:00Z" "1980-02-12T00:00:00Z" "1983-0
##  $ age_at_delivery: num  5.02 7.05 7.87 4.9 5.98 ...
##  $ age_at_death   : num  0.863 0.9123 0 0.0192 1.0219 ...
##  $ treatment      : chr  "control" "control" "control" "control" ...
```

```
# mean age at death of females
m_fem <- mean(fem$age_at_death)
m_fem
```

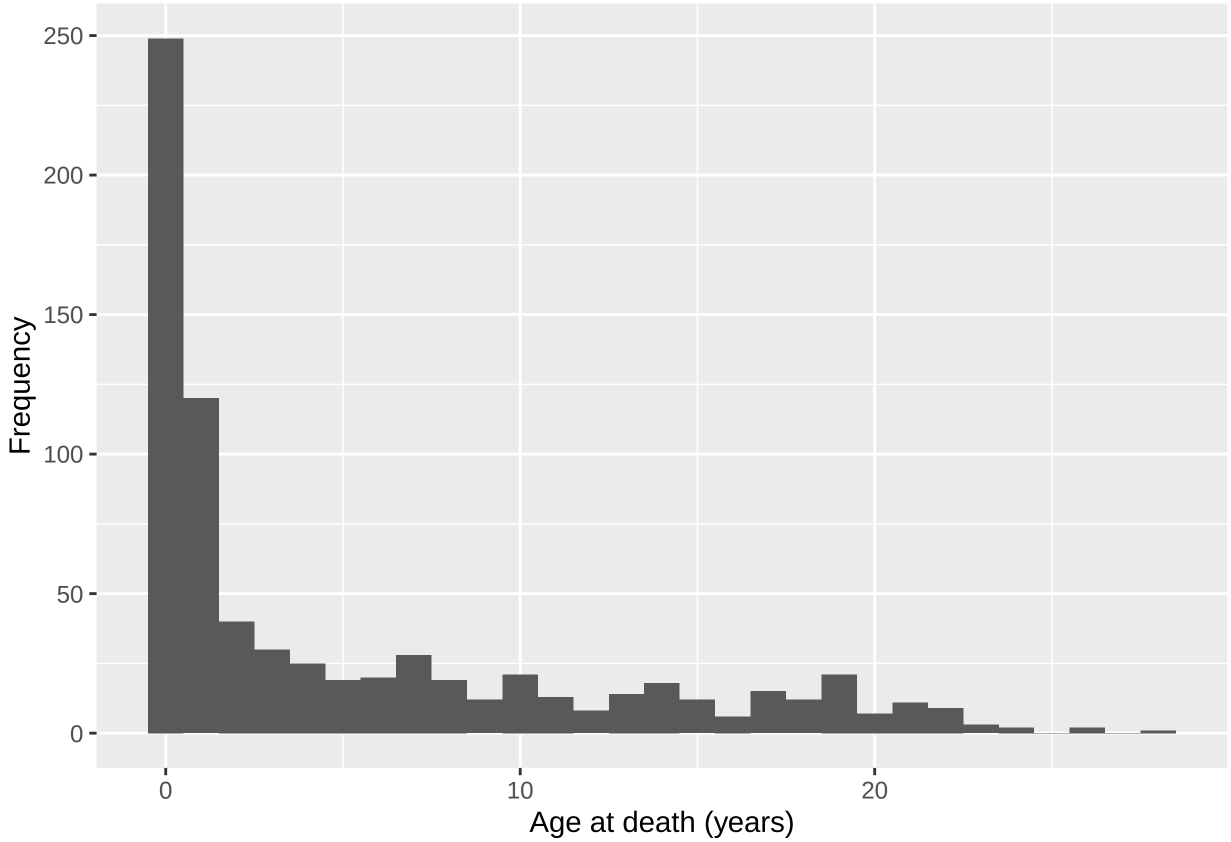
```
## [1] 5.641832
```

```
# standard deviation for females
s_fem <- sd(fem$age_at_death)
s_fem
```

```
## [1] 6.803812
```

Challenge 1: Plot the Distribution of Age at Death for Females.

```
# plotting counts
p3 <- ggplot(fem,aes(x=floor(age_at_death))) +
  geom_histogram(binwidth = 1) +
  xlab("Age at death (years)") +
  ylab("Frequency")
p3
```



Question Answers

a. Females in the age classes of 0 and 1 experience more deaths.
b. The pattern of age at death for females does not differ much at all from the entire population distribution.

Section 3 - Estimating the Sampling Distribution

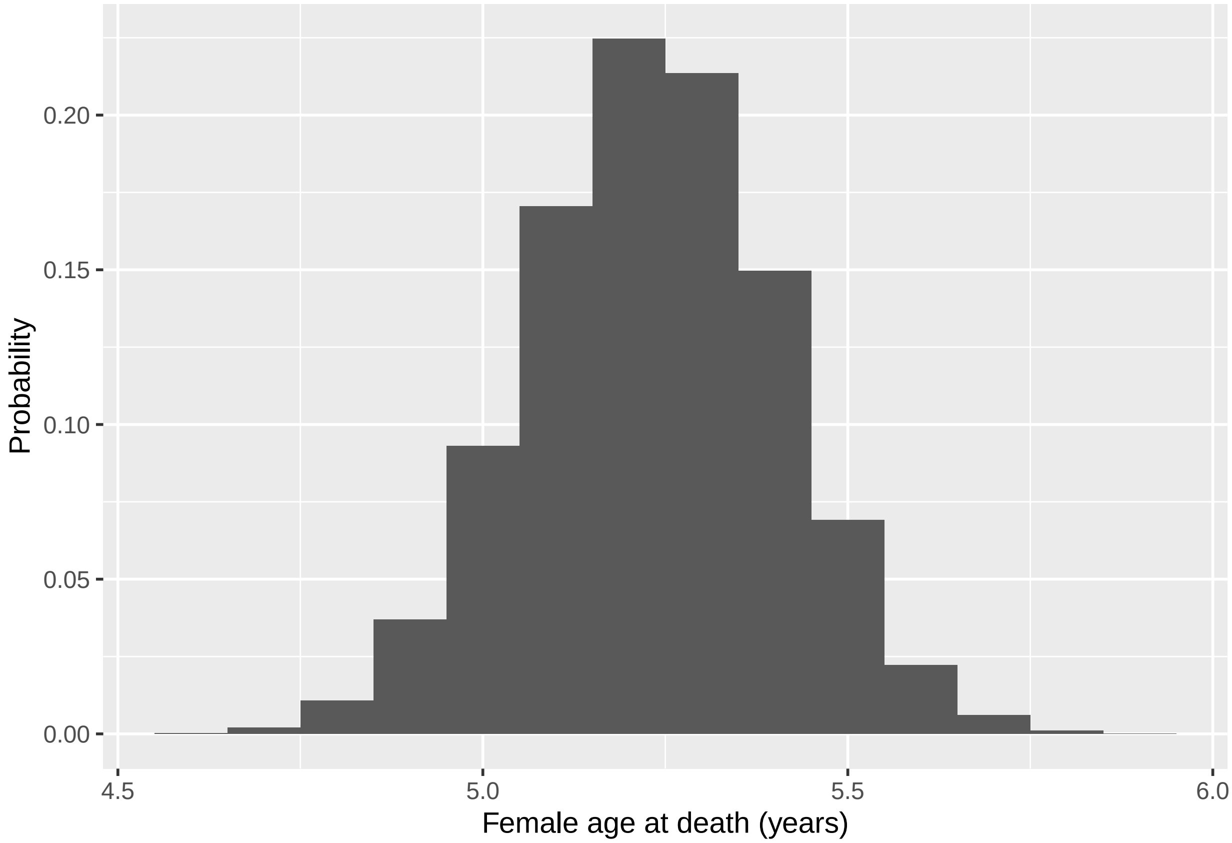
```r
# sample() allows to randomly sample a dataframe
# creating a new object to add the 10,000 sample statistics
r500 <- NULL

# loop for sampling 500 females at random and estimate the mean age at death
for(i in 1:10000){
  temporarySample <- sample(floor(fem$age_at_death), size = 500,
                                  replace = FALSE)
  r500[i] <- mean(temporarySample)
}

r500 <- as.data.frame(r500)

# viewing the new data frame created
#View(r500)
```

```
# plotting the sampling distribution
p4 <- ggplot(r500,aes(x=r500)) +
  geom_histogram(aes(y=..count../sum(..count..)),binwidth = .1) +
  xlab("Female age at death (years)") +
  ylab("Probability")
p4
```



Question Answers

a. The above sampling distribution is centered around the previous sample mean of 5.64 years of age at death for females.
b. The standard deviation value of 6.80 years of age at death for females in the previous sample serves as a proxy for the width of the normal distribution of 10000 means.
c. If we were to re-run the exercise with a sample of 50 females, the sample mean might be similar, but the sample standard deviation will increase.

Section 4 - Estimating the Standard Error

```
# standard error of the mean
se_fem <- s_fem/sqrt(737)
se_fem
```

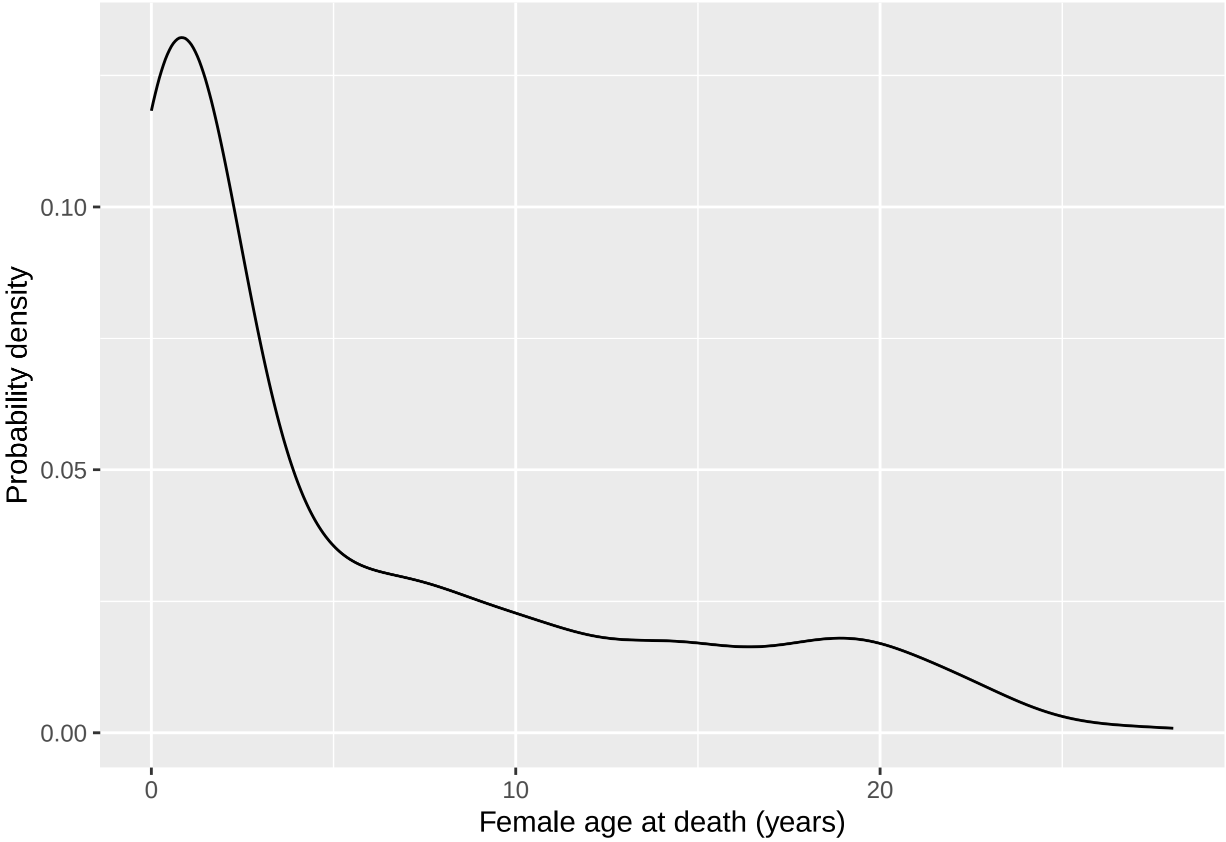
```
## [1] 0.2506216
```

Question Answer

a. With a Standard Error (SE) of 0.25 years, the sample mean for female monkeys is relatively precise.

Section 5 - Plotting the Probability Distribution of the Mean Age at Death

```
# probability density distribution for females
#geom_density() creates a smooth version of a histogram via kernel interpolation
p5 <- ggplot(fem,aes(age_at_death)) +
  geom_density() +
  xlab("Female age at death (years)") +
  ylab("Probability density")
p5
```



Question Answer

a. There is a pattern of much greater probabilities of death in the early years of female monkey lives, particularly just after birth. This makes sense biologically since newborn primates are very fragile and require a lot of care.

Section 6 - Estimating the Probability of Two or More (Survival) Events

```
# Dying and surviving in a particular age are two mutually exclusive events; we either die or survive.
# frequency table of age at death of females
table(floor(fem$age_at_death))
```

```
## 
##   0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19 
## 249 120  40  30  25  19  20  28  19  12  21  13   8  14  18  12   6  15  12  21 
##  20  21  22  23  24  26  28 
##   7  11   9   3   2   2   1
```

```
# proportion of deaths during the first year of life
m0 <- 249/737
m0
```

```
## [1] 0.3378562
```

```
# survival during the first year of life
s0 <- 1-m0
s0
```

```
## [1] 0.6621438
```

```
# proportion of deaths during the second year of life
m1 <- 120/(737-249)
m1
```

```
## [1] 0.2459016
```

```
# survival during the second year of life
s1 <- 1-m1
s1
```

```
## [1] 0.7540984
```

```
# probability of surviving from birth to 2 years of age
s0*s1
```

```
## [1] 0.4993216
```

Question Answers

a. Because I wanted to know the probability of surviving past the first two age classes, the probabilities of surviving past each stage must be multiplied.
b. The probability that a female monkey will survive to age 2 is 0.499.

Stop, Think, Do - Answer the Following Research Question: Do hurricanes affect survival from birth to 2 years of age?

```r
# Hurricane year dataframe for females
# filtering previously created fem dataframe by hurricane year
hurri <- filter(fem, treatment=="hurricane")
str(hurri)
```

```
## 'data.frame':    52 obs. of  8 variables:
##  $ dob            : chr  "1988-12-11T00:00:00Z" "1989-01-22T00:00:00Z" "1989-01-25T00:00:00Z" "1989-
##  $ season         : int  1989 1989 1989 1989 1986 1992 1998 1988 1997 1997 ...
##  $ sex            : chr  "f" "f" "f" "f" ...
##  $ dod            : chr  "1988-12-11T00:00:00Z" "1989-03-10T00:00:00Z" "1989-05-31T00:00:00Z" "1989-
##  $ mom_dob        : chr  "1982-02-04T00:00:00Z" "1985-02-28T00:00:00Z" "1983-02-07T00:00:00Z" "1981-
##  $ age_at_delivery: num  6.85 3.9 5.97 7.98 5.96 ...
##  $ age_at_death   : num  0 0.129 0.345 0.405 3.852 ...
##  $ treatment      : chr  "hurricane" "hurricane" "hurricane" "hurricane" ...
```

```r
# mean age at death of females
m_hurri <- mean(hurri$age_at_death)
m_hurri
```

```
## [1] 7.545522
```

```r
# standard deviation for females
s_hurri <- sd(hurri$age_at_death)
s_hurri
```

```
## [1] 8.364504
```

```r
# frequency table of age at death of females during hurricanes
table(floor(hurri$age_at_death))
```

```
## 
##  0  1  3  4  5  8  9 10 14 15 16 17 18 19 20 22 24 26 
## 16 10  3  1  2  1  2  2  1  1  1  3  3  1  1  2  1  1
```

```r
# Hurricane proportions
# proportion of deaths during the first year of life
m0_hurri <- 16/52
m0_hurri
```

```
## [1] 0.3076923
```

```r
# survival during the first year of life
s0_hurri <- 1-m0_hurri
s0_hurri
```

```
## [1] 0.6923077
```

```r
# proportion of deaths during the second year of life
m1_hurri <- 10/(52-16)
m1_hurri
```

```
## [1] 0.2777778
```

```r
# survival during the second year of life
s1_hurri <- 1-m1_hurri
s1_hurri
```

```
## [1] 0.7222222
```

```r
# probability of surviving from birth to 2 years of age
s0_hurri*s1_hurri
```

```
## [1] 0.5
```

```r
# Normal or control year dataframe for females
# filtering previously created fem dataframe by normal or control year
ctrl <- filter(fem, treatment=="control")
str(ctrl)
```

```
## 'data.frame':    685 obs. of  8 variables:
##  $ dob             : chr  "1987-03-02T00:00:00Z" "1987-02-21T00:00:00Z" "1987-12-24T00:00:00Z" "1988-
##  $ season          : int  1987 1987 1988 1988 1987 1988 1985 1990 1990 1990 ...
##  $ sex             : chr  "f" "f" "f" "f" ...
##  $ dod             : chr  "1988-01-11T00:00:00Z" "1988-01-20T00:00:00Z" "1987-12-24T00:00:00Z" "1988-
##  $ mom_dob         : chr  "1982-02-22T00:00:00Z" "1980-02-04T00:00:00Z" "1980-02-12T00:00:00Z" "1983-
##  $ age_at_delivery: num  5.02 7.05 7.87 4.9 5.98 ...
##  $ age_at_death   : num  0.863 0.9123 0 0.0192 1.0219 ...
##  $ treatment      : chr  "control" "control" "control" "control" ...
```

```r
# mean age at death of females
m_ctrl <- mean(ctrl$age_at_death)
m_ctrl
```

```
## [1] 5.497318
```

```r
# standard deviation for females
s_ctrl <- sd(ctrl$age_at_death)
s_ctrl
```

```
## [1] 6.655676
```

```r
# frequency table of age at death of females during normal years
table(floor(ctrl$age_at_death))
```

```
## 
##   0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19 
## 233 110  40  27  24  17  20  28  18  10  19  13   8  14  17  11   5  12   9  20 
##  20  21  22  23  24  26  28 
##   6  11   7   3   1   1   1 
```

```r
# Hurricane proportions
# proportion of deaths during the first year of life
m0_ctrl <- 233/685
m0_ctrl
```

```
## [1] 0.340146
```

```r
# survival during the first year of life
s0_ctrl <- 1-m0_ctrl
s0_ctrl
```

```
## [1] 0.659854
```

```r
# proportion of deaths during the second year of life
m1_ctrl <- 110/(685-233)
m1_ctrl
```

```
## [1] 0.2433628
```

```r
# survival during the second year of life
s1_ctrl <- 1-m1_ctrl
s1_ctrl
```

```
## [1] 0.7566372
```

```r
# probability of surviving from birth to 2 years of age
s0_ctrl*s1_ctrl
```

```
## [1] 0.4992701
```

The above analysis demonstrates that the probability of female survival to two years of age does not differ between hurricane years (0.500) and normal years (0.499).

Discussion Question Answers

a. A population distribution would be the standard length measurements of all California halibut in San Francisco Bay on the first day of May, 2025. A sampling distribution would be the frequency of means of multiple random samples of 100 captured and measured California Halibut in San Francisco Bay.
b. The measure of precision in the previous lab (Chapter 2) involved taking the range of measured values for a sample size of three observations. The measure of precision in this lab involved calculating the standard error of the mean for a sample size of 737 observations.
c. To apply the multiplication rule in our survival exercise, we must assume that death and survival in different age classes are mutually exclusive events.