Lab 4

Adam Orr

September 20, 2017

Section: 91973 Friday 9am

Question 1

Script

```
#import the data
setwd("~/code/biometry-lab/lab4")
df <- read.csv('lizard.csv')
(lengths <- df$length)

#estimate the mean of length by direct calculation
(xbar <- sum(lengths)/length(lengths))

#estimate the mean with R's built in function
mean(lengths)

#estimate the variance of length by direct calculation
(s2 <- sum((lengths  - xbar)^2)/(length(lengths)-1))

#estimate the variance with R's built in function
var(lengths)

#estimate the standard deviation of length by direct calculation
(s <- sqrt(s2))

#estimate the standard deviation with R's built-in function
sd(lengths)

#estimate the standard error of the mean length
(sem <- s / sqrt(length(lengths)))

#and calculate the approximate 95% confidence interval
c(xbar - 2*sem, xbar + 2 * sem)
```

Output

```
#import the data
```

```
##  [1]  82  94  93  88 105  76  78  99 113  78 148 100  65  99  94
```

```
#estimate the mean of length by direct calculation
```

```
## [1] 94.13333
```

```
#estimate the mean with R's built in function
```

```
## [1] 94.13333
```

```
#estimate the variance of length by direct calculation
```

```
## [1] 381.5524
```

```
#estimate the variance with R's built in function
```

```
## [1] 381.5524
```

```
#estimate the standard deviation of length by direct calculation
```

```
## [1] 19.53337
```

```
#estimate the standard deviation with R's built-in function
```

```
## [1] 19.53337
```

```
#estimate the standard error of the mean length
```

```
## [1] 5.043493
```

```
#and calculate the approximate 95% confidence interval
```

```
## [1]  84.04635 104.22032
```

Answers

Is the estimated mean likely to equal the population parameter? Why or why not?

The estimated mean is not likely to be exactly equal to the population parameter because the sample is chosen randomly. However, the estimated mean is likely to be somewhat close to the true population parameter.

Which is a better descriptor of the variation in length of lizards, the standard deviation or the standard error? Why?

The variation of the lizard lengths is better described by the standard deviation because it is a description of the variability in our sample. The standard error is a description of the variability of our estimate of the mean.

Which is a better descriptor of the uncertainty in the estimated mean length of lizards? Why?

The standard error is a better descriptor of the uncertainty in the estimate of the mean because it describes the standard deviation of the distribution of $\bar{X}$. On the other hand, the standard deviation describes the variability in our sample of lizard lengths.

Question 2

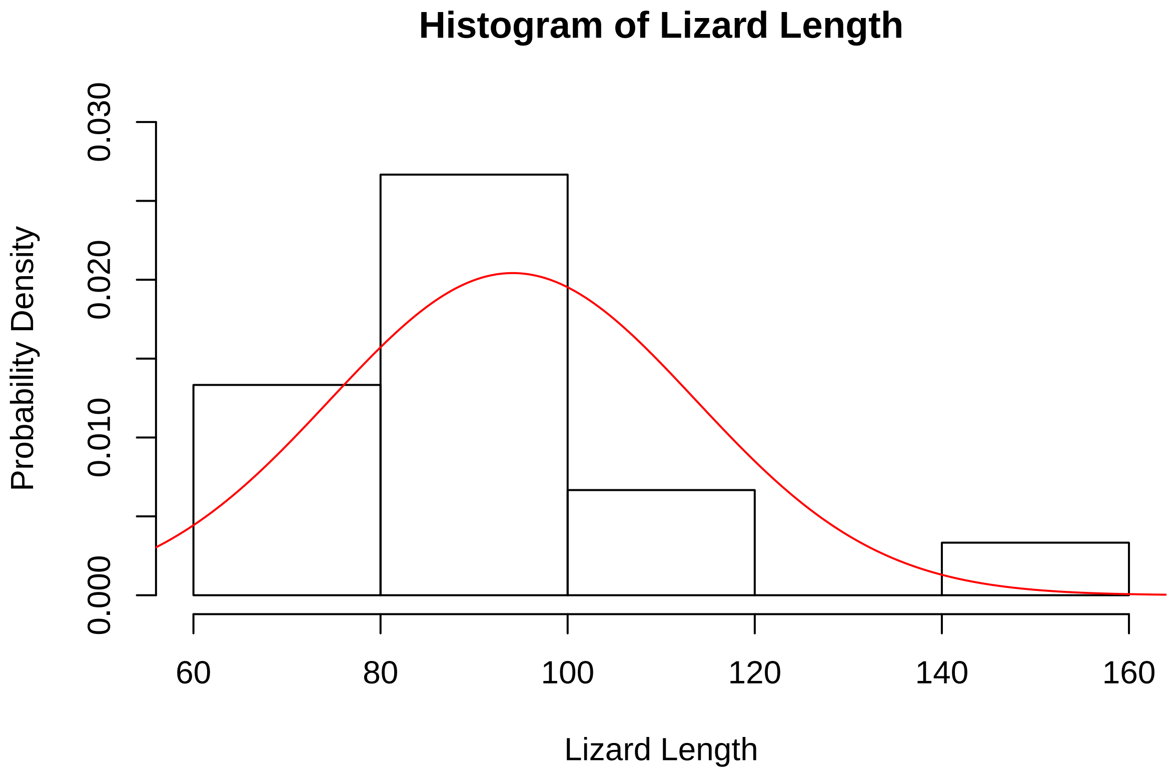
Script

```
#make a histogram of the lizard lengths
hist(lengths, xlab = "Lizard Length",
     ylab = "Probability Density",
     main = "Histogram of Lizard Length",
     freq = FALSE, ylim = c(0,.03))

#add a curve to the histogram showing your estimate of the pdf
x<-seq(xbar-4*s, xbar+4*s, by = .1)
pdf <- dnorm(x,xbar,s)
lines(pdf~x, col='red')
```

Output

```
#make a histogram of the lizard lengths
#add a curve to the histogram showing your estimate of the pdf
```



Answers

Question 3

Script

```
#redraw plot
hist(lengths, xlab = "Lizard Length",
     ylab = "Probability Density",
     main = "Histogram of Lizard Length",
     freq = FALSE, ylim = c(0,.12))
lines(pdf~x, col = 'red')

#add another curve to the histogram showing the pdf of the average of 15 lizard lengths
```

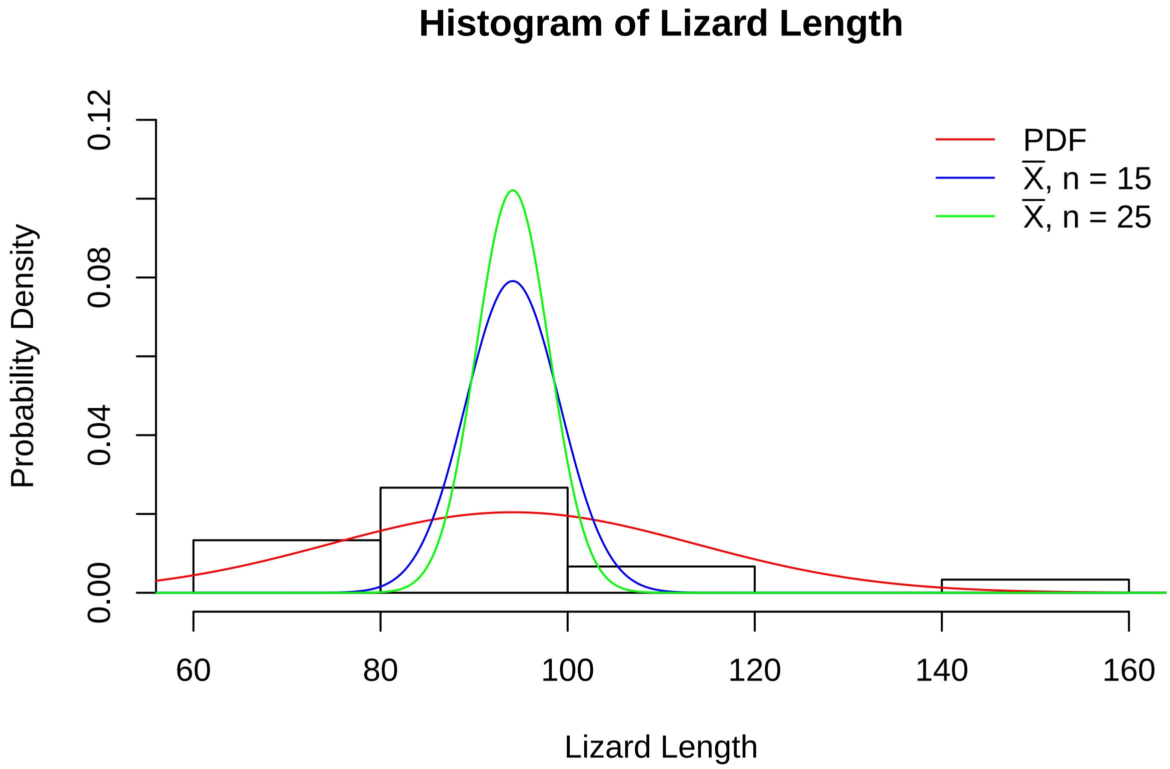
```r
pdf15 <- dnorm(x,xbar, s / sqrt(15))
lines(pdf15~x, col = 'blue')

#add one more curve showing the pdf of the average of 25 lizard lengths
pdf25 <- dnorm(x,xbar, s / sqrt(25))
lines(pdf25~x, col = 'green')

#add a legend
legend("topright",
       legend = c("PDF", expression(paste(bar(X), ", n = 15")), expression(paste(bar(X),", n = 25"))),
       col = c('red','blue','green'),
       lty = 1,
       bty = "n")
```

Output

```r
#redraw plot
#add another curve to the histogram showing the pdf of the average of 15 lizard lengths
#add one more curve showing the pdf of the average of 25 lizard lengths
#add a legend
```



Answers

Compare the location and dispersion of the distributions of the three variables: length, average length for a sample of 15, average length for a sample of 25. Interpret the pattern you see to explain why the average is an accurate estimator of the mean, and how sample size effects the quality of the estimate.

The distributions all have the same location. However, as the sample size increases, the dispersion of the estimate of the mean decreases. Since the distribution of the average has a mean value of the mean of the population, it is an accurate estimator of the mean. This estimate will be more accurate with increasing sample size, as the dispersion of the estimate decreases as the amount of data increases.

Question 4

Script

```
#actual population parameters
mu <- 12
sigma <- 2

#what is the probability that a randomly sampled male will have a horn less than 11 mm long?
pnorm(11,mu,sigma)

#What is the probability that the average length of a random sample of 10 horns will be less than 11 mm
pnorm(11,mu,sigma/sqrt(10))

#What is the probability that the average length of a random sample of 10 horns will be greater than 12
1 - pnorm(12.5, mu, sigma/sqrt(10))

#What is the probability that the average length of a random sample of 50 horns will be greater than 12
1 - pnorm(12.5, mu, sigma/sqrt(50))

#What are the 0.025 and 0.975 quantiles of average horn length, for a sample of 10?
#0.025 quantile
qnorm(.025, mu, sigma/sqrt(10))

#0.975 quantile
qnorm(.975, mu, sigma/sqrt(10))
```

Output

```
#actual population parameters
#what is the probability that a randomly sampled male will have a horn less than 11 mm long?
```

```
## [1] 0.3085375
```

```
#What is the probability that the average length of a random sample of 10 horns will be less than 11 mm
```

```
## [1] 0.05692315
```

```
#What is the probability that the average length of a random sample of 10 horns will be greater than 12
```

```
## [1] 0.2145977
```

```
#What is the probability that the average length of a random sample of 50 horns will be greater than 12
```

```
## [1] 0.03854994
```

```
#What are the 0.025 and 0.975 quantiles of average horn length, for a sample of 10?
#0.025 quantile
```

```
## [1] 10.76041
```

```
#0.975 quantile
```

```
## [1] 13.23959
```

Answers

Explain the difference in the probabilities you calculated.

The first probability describes how often a randomly sampled male will have a horn less than 11mm long, while the second probability describes how often an average of a sample of 10 males is less than 11mm.

Explain the difference in the probabilities you calculated.

The first probability describes how often the average of a random sample of 10 males will be larger than 12.5 mm, while the second probability describes how often the average of a random sample of 50 males will be larger than 12.5 mm.

Question 5

Script

```
mu <- 6.2
sigma2 <- 0.25
sigma <- sqrt(sigma2)

#Generate a sample of 10 randomly samples cone weights.
(sample <- rnorm(10,mu,sigma))

#Use your sample to estimate the mean, variance, and standard deviation using R's built-in functions
#mean
mean(sample)
#variance
var(sample)
#standard deviation
sd(sample)

#Use your sample to estimate the standard error of the mean
(sem <- sd(sample)/sqrt(length(sample)))

#and calculate the approximate 95% confidence interval
c(mean(sample) - 2 * sem, mean(sample) + 2 * sem)
```

Output

```
#Generate a sample of 10 randomly samples cone weights.
```

```
##  [1] 6.437186 5.890413 5.954342 5.561659 5.600925 5.752165 5.449755
##  [8] 6.490771 6.136141 6.205068
```

```
#Use your sample to estimate the mean, variance, and standard deviation using R's built-in functions
#mean
```

```
## [1] 5.947842
```

```
#variance
```

```
## [1] 0.1327845
```

```
#standard deviation
```

```
## [1] 0.3643961
```

```
#Use your sample to estimate the standard error of the mean
```

```
## [1] 0.1152322
```

```
#and calculate the approximate 95% confidence interval
```

```
## [1] 5.717378 6.178307
```

Answers

Run the script again. Are the boundaries of the confidence interval the same as before? Why or why not?

No, the boundaries of the confidence interval are not the same as before. This is because the data used to generate the estimates is obtained randomly.

Did your confidence interval include the true value of the mean? If you repeated this exercise 100 times, roughly how many times would you expect the interval to include the true mean?

Yes, the confidence interval included the true value of the mean. If I repeated this 100 times, I would expect 95 times would include the true mean.

Now run your script once again, but first make one change: use a sample size of 90 instead of 10. Compare the new confidence interval to the first one you generated. What happened to the interval when you changed sample size, and why?

The confidence interval decreased in size. This happened because a larger sample size reduces the standard error of the mean, decreasing the width of the confidence interval.

Assuming that you did not already know the true value of the mean, which of your two estimates would you trust more, and why?

I would trust the estimate that was generated using a larger sample size. This is because there is less variance in the distribution of estimates generated using larger sample sizes, so it is more likely to be closer to the true mean.

Question 6

Script

Output

Answers

State the expected effect of increasing sample size. All these effects were observed in my own calculations in exercise 5.

Estimate of the mean weight.

The estimate of the mean weight would not change.

Estimate of the variance of weight.

The estimate of the variance of weight would not change.

The estimated standard error of the mean weight.

The estimate of the standard error of the mean weight would decrease.

The width of the 95% confidence interval.

The width of the 95% condifence interval would decrease.