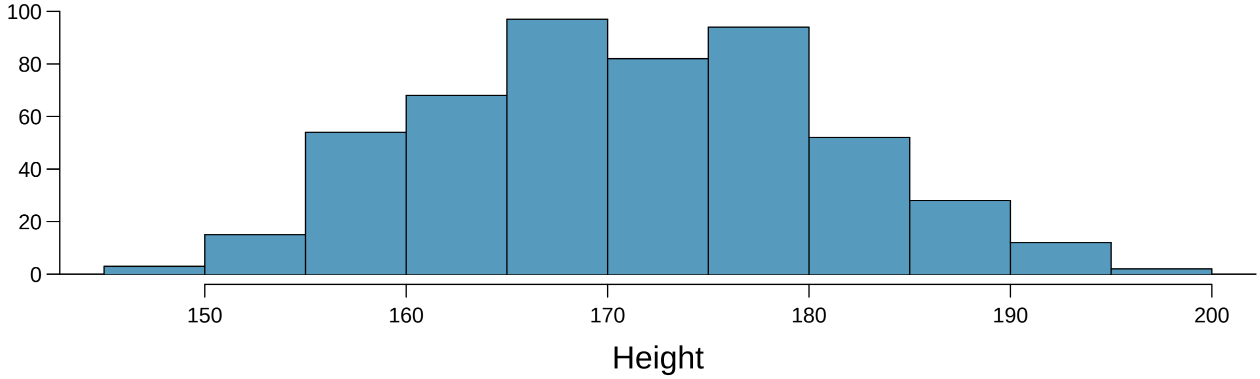
Chapter 5 - Foundations for Inference

Subhalaxmi Rout

Heights of adults. (7.7, p. 260) Researchers studying anthropometry collected body girth measurements and skeletal diameter measurements, as well as age, weight, height and gender, for 507 physically active individuals. The histogram below shows the sample distribution of heights in centimeters.



(a) What is the point estimate for the average height of active individuals? What about the median?
(b) What is the point estimate for the standard deviation of the heights of active individuals? What about the IQR?
(c) Is a person who is 1m 80cm (180 cm) tall considered unusually tall? And is a person who is 1m 55cm (155cm) considered unusually short? Explain your reasoning.
(d) The researchers take another random sample of physically active individuals. Would you expect the mean and the standard deviation of this new sample to be the ones given above? Explain your reasoning.
(e) The sample means obtained are point estimates for the mean height of all active individuals, if the sample of individuals is equivalent to a simple random sample. What measure do we use to quantify the variability of such an estimate (Hint: recall that $SD_x = \frac{\sigma}{\sqrt{n}}$)? Compute this quantity using the data from the original sample under the condition that the data are a simple random sample.

1. Heights of adults

(a) Point estimate = 171.1
Median = 170.3

(b) Standard Deviation = 9.4
IQR = Q3 - Q1 = 177.8 - 163. 8 = 14

(c) A person who is 1m 80cm (180 cm) tall.

```r
x = 180
mean = 171.1
sd = 9.4
 z <- (x - mean) / sd
 paste0("value of z is ", round(z,2))
```

```
## [1] "value of z is 0.95"
```

A person who is 1m 55cm (155 cm) tall.

```
x = 155
mean = 171.1
sd = 9.4
 z <- (x - mean) / sd
 paste0("value of z is ", round(z,2))
```

```
## [1] "value of z is -1.71"
```

From Z value we can consider that a person whose height is 180 near is 0.95 Standard Deviation away from the mean which is not that unusual. The person whose height is 155, he is 1.71 Standard Deviation away from the mean which is more unusual.

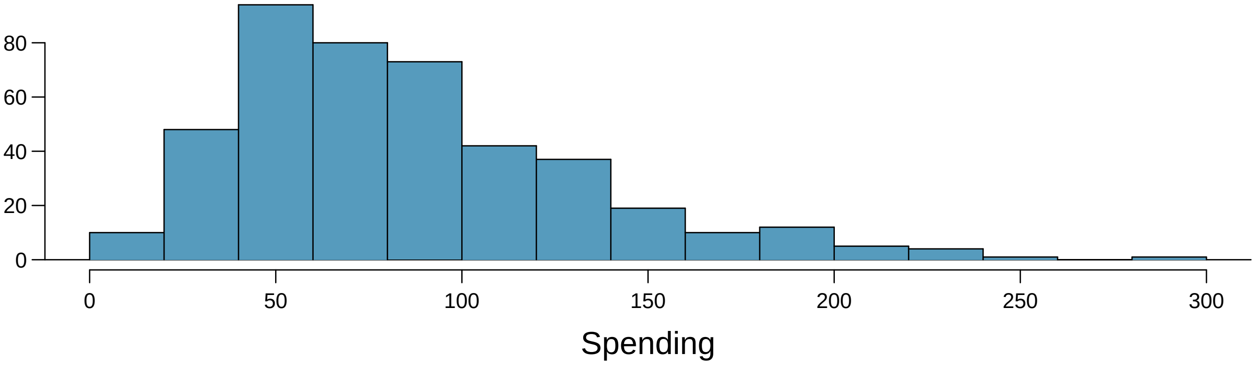
(d) I think mean and Standard Deviation would not be the same beacuse this would be a different set of sample data so mean and variability can varies according to the population.

(e) $SD_x = \frac{\sigma}{\sqrt{n}}$

```
sigma =  171.1
n = 507
Stand_error <- sigma / sqrt(n)
paste0("SE (standard error) is calculate variability of sample which is ", round(Stand_error,2))
```

```
## [1] "SE (standard error) is calculate variability of sample which is 7.6"
```

Thanksgiving spending, Part I. The 2009 holiday retail season, which kicked off on November 27, 2009 (the day after Thanksgiving), had been marked by somewhat lower self-reported consumer spending than was seen during the comparable period in 2008. To get an estimate of consumer spending, 436 randomly sampled American adults were surveyed. Daily consumer spending for the six-day period after Thanksgiving, spanning the Black Friday weekend and Cyber Monday, averaged \$84.71. A 95% confidence interval based on this sample is (\$80.31, \$89.11). Determine whether the following statements are true or false, and explain your reasoning.



(a) We are 95% confident that the average spending of these 436 American adults is between \$80.31 and \$89.11.
(b) This confidence interval is not valid since the distribution of spending in the sample is right skewed.
(c) 95% of random samples have a sample mean between \$80.31 and \$89.11.
(d) We are 95% confident that the average spending of all American adults is between \$80.31 and \$89.11.
(e) A 90% confidence interval would be narrower than the 95% confidence interval since we don't need to be as sure about our estimate.
(f) In order to decrease the margin of error of a 95% confidence interval to a third of what it is now, we would need to use a sample 3 times larger.
(g) The margin of error is 4.4.

2. Thanksgiving spending, Part I

(*a*) FALSE: Confidence Interval (CI) applies to the population not the sample.

(*b*) FALSE: The confidence interval may still valid when the sample distribution is skewed.

(*c*) FALSE: Samples of different size may have different confidence interval.

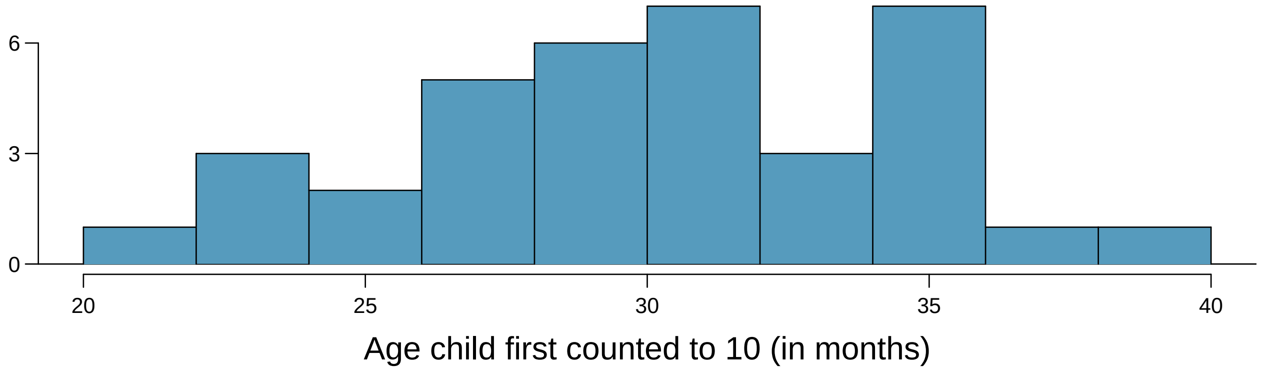
(*d*) TRUE: Confidence Interval covers parameter value.

(*e*) TRUE: Statement is correct.

(*f*) FALSE: In order to decraese the margin error of a 95% confidence interval to a third, we need a sample size (3 ˆ 2 = 9) 9 times lager.

(*g*) TRUE: Margin of error: (89.11-80.31)/2 = 4.4

Gifted children, Part I. Researchers investigating characteristics of gifted children col- lected data from schools in a large city on a random sample of thirty-six children who were identified as gifted children soon after they reached the age of four. The following histogram shows the dis- tribution of the ages (in months) at which these children first counted to 10 successfully. Also provided are some sample statistics.



n	36
min	21
mean	30.69
sd	4.31
max	39

(a) Are conditions for inference satisfied?
(b) Suppose you read online that children first count to 10 successfully when they are 32 months old, on average. Perform a hypothesis test to evaluate if these data provide convincing evidence that the average age at which gifted children fist count to 10 successfully is less than the general average of 32 months. Use a significance level of 0.10.
(c) Interpret the p-value in context of the hypothesis test and the data.
(d) Calculate a 90% confidence interval for the average age at which gifted children first count to 10 successfully.
(e) Do your results from the hypothesis test and the confidence interval agree? Explain.

3. Gifted children, Part I

(a) Yes.

- Children were randomly selected
- They are indepentdent on each other.
- Sample size is more than 30 (n = 36).

(b) Suppose Null Hypothesis (Hn)= 32 Alternate Hypothesis (Ha) < 32 significance level = 0.10

```
mean <- 30.69
n <-  36
sd <- 4.31
x <- 32
SE <- sd / sqrt(n)
z <- round((mean - x) / SE,2)
z
```

```
## [1] -1.82
```

```
p_value <- round(pnorm(z, mean = 0, sd = 1),2)
p_value
```

```
## [1] 0.03
```

p_value = 0.03 significance level = 0.10 So, the p-value < the significance level, reject the null hypothesis.

(*c*) Since the p-value (0.03) is less than the significance level of .1, we reject the null hypothesis in the favour of alternative hypothesis.

(*d*)

```
mean <- 30.69
n <-  36
sd <- 4.31
SE <- sd / sqrt(n)
# for .90 CI, value of z is 1.645
z <- 1.645
high <- round(mean + (z * SE),2)
paste0("Higher age value ", high)
```

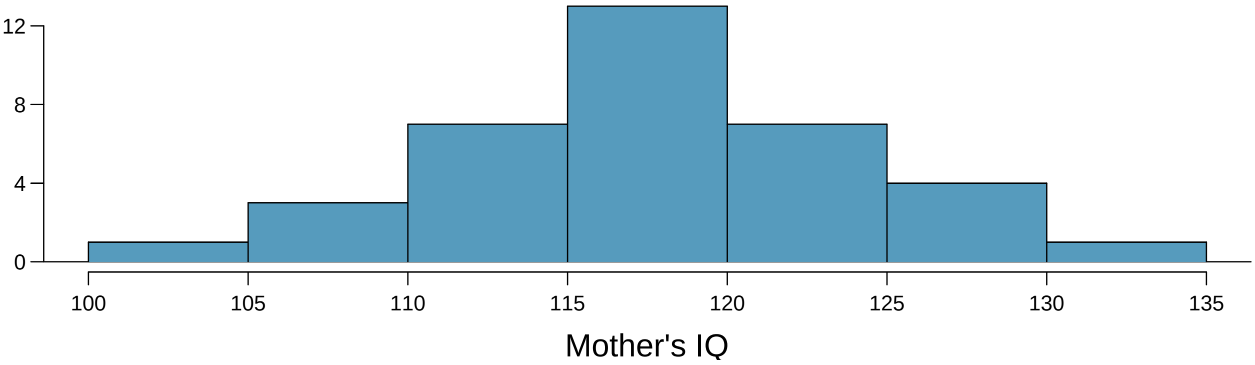
```
## [1] "Higher age value 31.87"
```

```
low <- round(mean - (z * SE),2)
paste0("Lower age value ", low)
```

```
## [1] "Lower age value 29.51"
```

(*e*) Yes, they are. The confidence interval says that 90% chance the true mean for gifted children is between 29.51 and 31.87 months.

Gifted children, Part II. Exercise above describes a study on gifted children. In this study, along with variables on the children, the researchers also collected data on the mother's and father's IQ of the 36 randomly sampled gifted children. The histogram below shows the distribution of mother's IQ. Also provided are some sample statistics.



n	36
min	101
mean	118.2
sd	6.5
max	131

(a) Perform a hypothesis test to evaluate if the sedata provide convincing evidence that the average IQ of mothers of gifted children is different than the average IQ for the population at large, which is 100. Use a significance level of 0.10.

(b) Calculate a 90% confidence interval for the average IQ of mothers of gifted children.

(c) Do your results from the hypothesis test and the confidence interval agree? Explain.

4. Gifted children, Part II

(*a*) Suppose Null Hypothesis (Hn) $= 100$ Alternate Hypothesis (Ha) > 100 significance level $= 0.10$

```
n <- 36
mean <- 118.2
sd <- 6.5
x <- 100
z <- round((mean - x) / sd,2)
z
```

```
## [1] 2.8
```

```
p_value <- 1 - round(pnorm(z, mean = 0, sd = 1),3)
p_value
```

```
## [1] 0.003
```

p value (0.003 ~ 0.0), which is smaller than significance level(0.1) so reject the null hypothesis that the average IQ of mothers of gifted children is different than the average IQ for the population at large.

(*b*)

```r
mean <- 118.2
n <-  36
sd <- 6.5
SE <- sd / sqrt(n)
# for .90 CI, value of z is 1.645
z <- 1.645
high <- round(mean + (z * SE),2)
paste0("Higher age value ", high)
```

```
## [1] "Higher age value 119.98"
```

```r
low <- round(mean - (z * SE),2)
paste0("Lower age value ", low)
```

```
## [1] "Lower age value 116.42"
```

(*c*) Since 100 lies out side of confidence interval so in this case reject the null hypothesis.

CLT. Define the term “sampling distribution” of the mean, and describe how the shape, center, and spread of the sampling distribution of the mean change as sample size increases.

5. CLT

Sampling distribution of the mean shows how means from multiple samples of a population are distributed.The sampling distribution looks similar to the normal distribution. As sample size increases:

- Shape: close to normal distribution
- Center: has heightest value
- Spread: becomes narrow

CFLBs. A manufacturer of compact fluorescent light bulbs advertises that the distribution of the lifespans of these light bulbs is nearly normal with a mean of 9,000 hours and a standard deviation of 1,000 hours.

(a) What is the probability that a randomly chosen light bulb lasts more than 10,500 hours?
(b) Describe the distribution of the mean lifespan of 15 light bulbs.
(c) What is the probability that the mean lifespan of 15 randomly chosen light bulbs is more than 10,500 hours?
(d) Sketch the two distributions (population and sampling) on the same scale.
(e) Could you estimate the probabilities from parts (a) and (c) if the lifespans of light bulbs had a skewed distribution?

6. CFLBs

(a)

```
mean <- 9000
sd <- 1000

probablity <- 1 - (pnorm(10500, mean, sd))
# convert in %, multiply with 100
Prob <- round(probablity * 100,2)
paste0("The probability that a randomly chosen light bulb lasts more than 10,500 hours is ",Prob,"%")
```

```
## [1] "The probability that a randomly chosen light bulb lasts more than 10,500 hours is 6.68%"
```

(b) Since the population has normal distribution N(mean = 9000, sd = 1000) so the sample also have approximate normal distribution which is N(mean = 9000, SE = 1000/sqrt(15)).

(c)

```
n <- 15
mean <- 9000
sd <- 1000
SE <- sd/sqrt(15)
x <- 10500
z <- (x - mean) / SE
probability <- round(1 -  pnorm(z),4)
paste0("The probability that the mean lifespan of 15 randomly chosen light bulbs is more than 10,500 hou
```

```
## [1] "The probability that the mean lifespan of 15 randomly chosen light bulbs is more than 10,500 hou
```

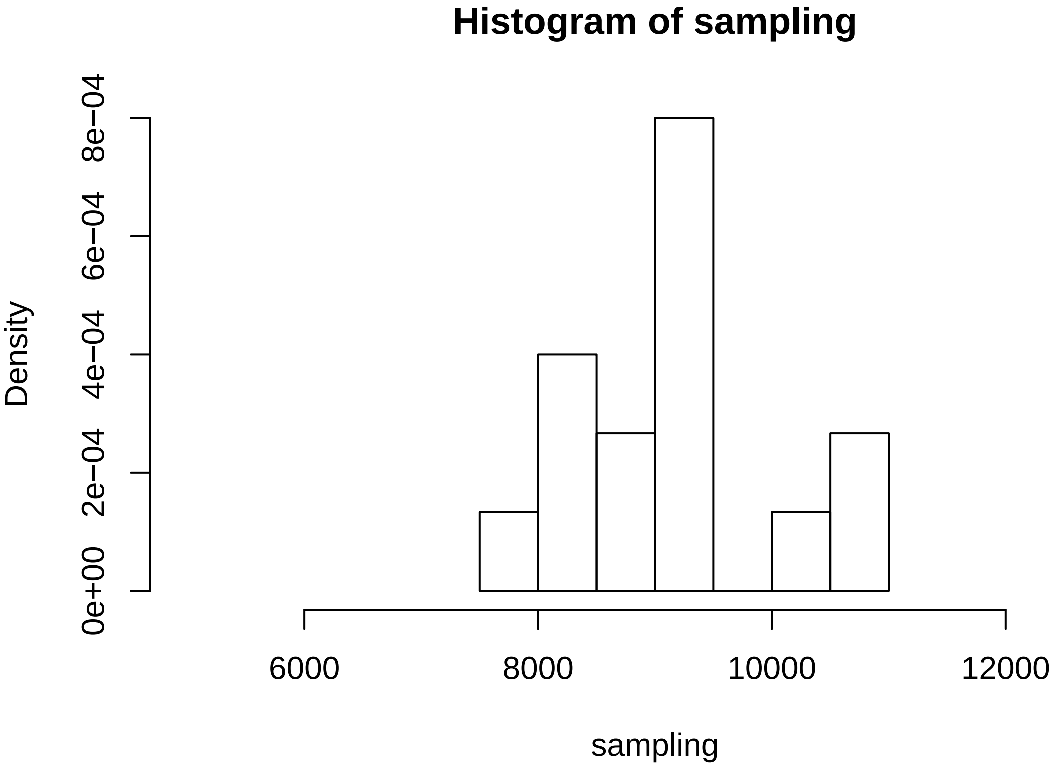
(d)

```
library(ggplot2)
```

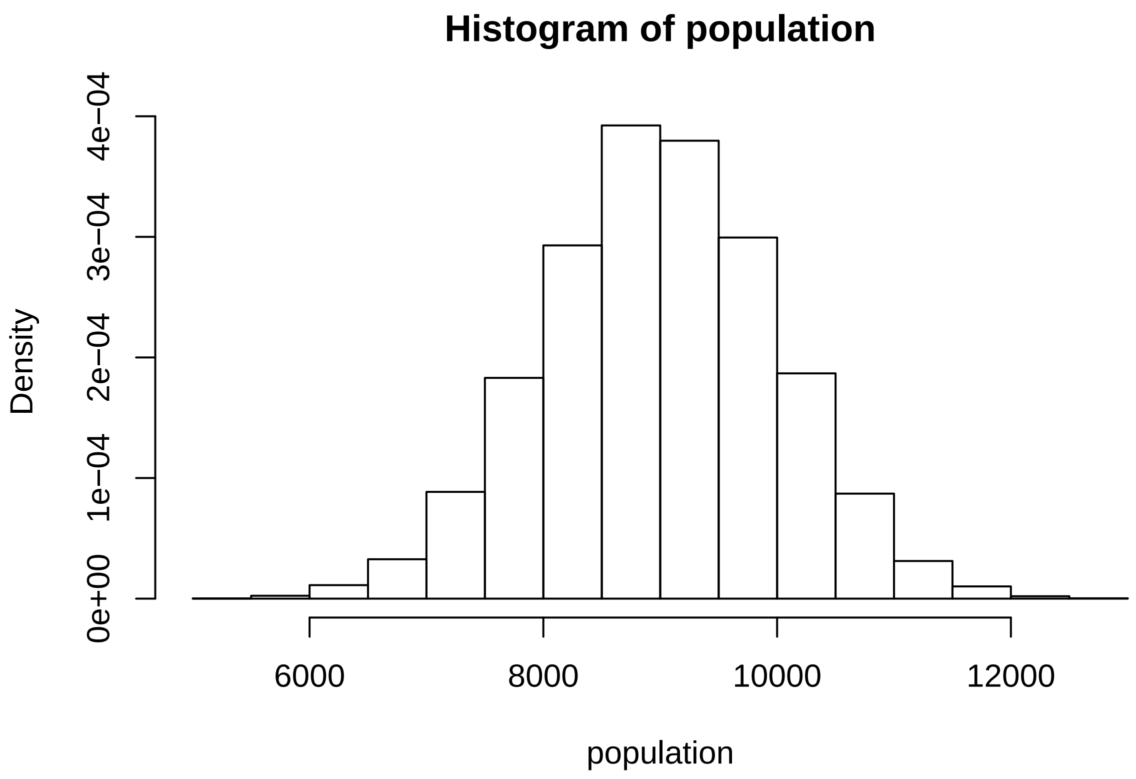
```
##
## Attaching package: 'ggplot2'

## The following object is masked from 'package:openintro':
##
##     diamonds
```

```
set.seed(123)
sampling <- rnorm(15, mean = 9000, sd = 1000)
population <- rnorm(15000, mean = 9000, sd = 1000)
hist(sampling, freq = FALSE, xlim=c(5000,13000))
```

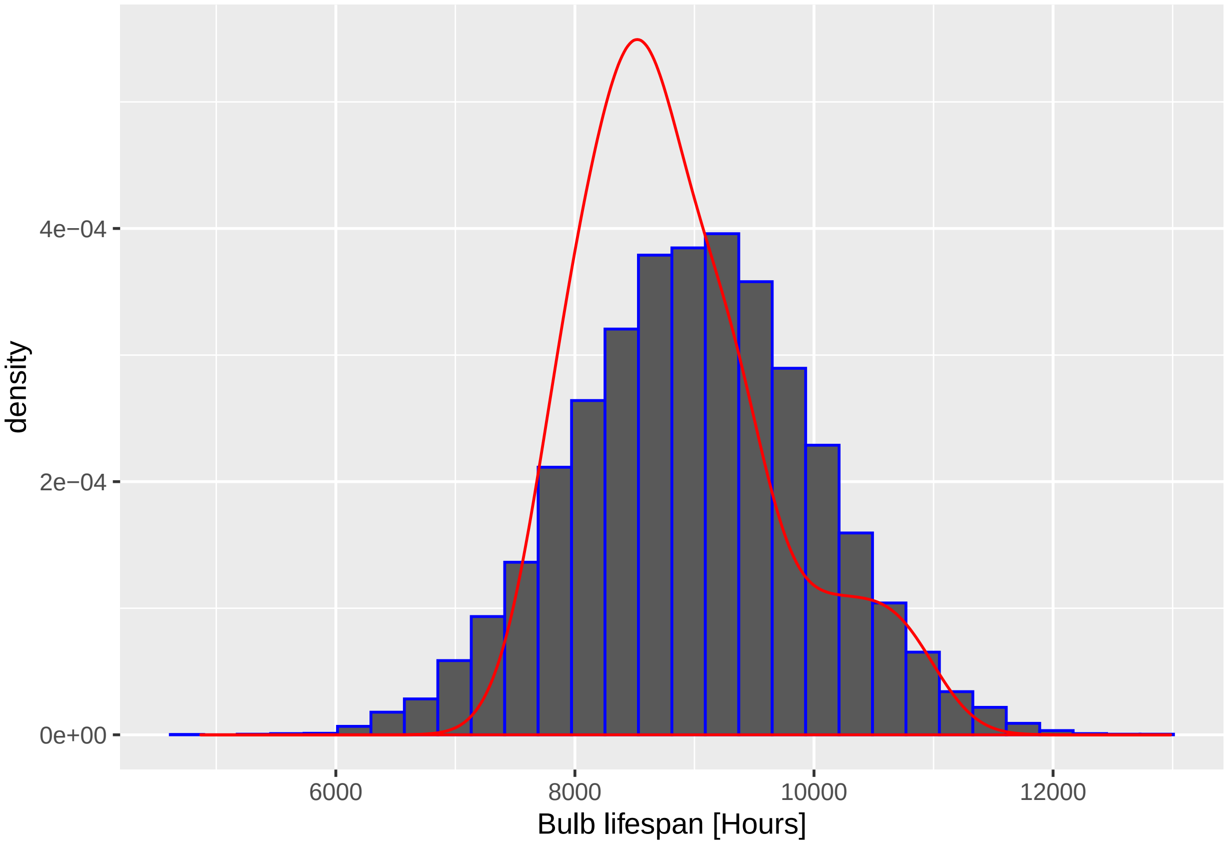


```
hist(population, freq = FALSE, xlim=c(5000,13000))
```



```
ggplot() +
geom_histogram(aes(x=rnorm(15000, mean = 9000, sd = 1000),y=..density..),color="blue") + xlab("Bulb lif
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



(e) No, we could not estimate the probability form (a) and (c) with skewed using normal distribution.

- In part (a), we would need to use the skewed distribution to do the probability calculations.
- In part (c), because n = 15 < 30, we need more sample size.

Same observation, different sample size. Suppose you conduct a hypothesis test based on a sample where the sample size is n = 50, and arrive at a p-value of 0.08. You then refer back to your notes and discover that you made a careless mistake, the sample size should have been n = 500. Will your p-value increase, decrease, or stay the same? Explain.

7. Same observation, different sample size

```r
sd <- 1
SE1 <- sd/sqrt(50) # case1 : n = 50
paste0("Standard error where n = 50 is ",round(SE1,3))
```

```
## [1] "Standard error where n = 50 is 0.141"
```

```r
SE2 <- sd/sqrt(500) # case2 : n = 500
paste0("Standard error where n = 500 is ",round(SE2,3))
```

```
## [1] "Standard error where n = 500 is 0.045"
```

For Z-score, SE (standard error) uses in the denominator. So, when the sample increases SE decreases, the Z-score will increase. If the Z-score increase the p-value will decrease. Above explanation, we got that, the p-value will decrease if the sample increases.