Homework Assigment 4

CUNY MSDA DATA 606

Duubar Villalobos Jimenez mydvtech@gmail.com

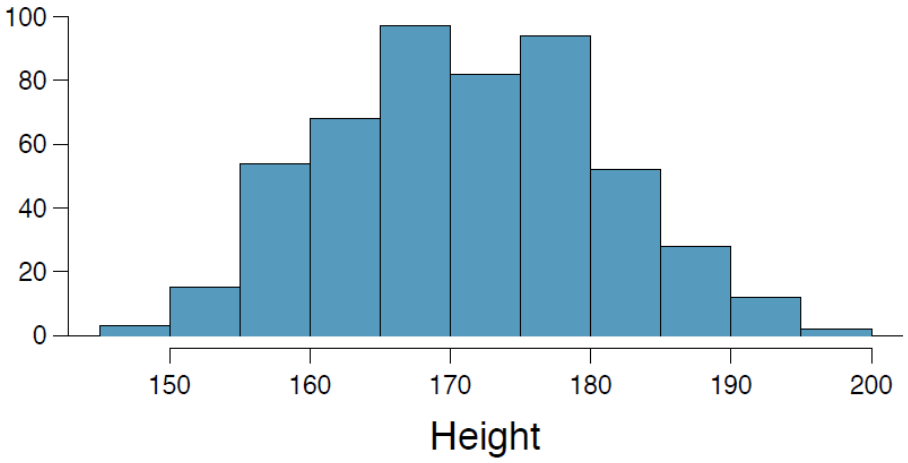
March 12, 2017

Chapter 4 Foundations for Inference

```
Practice: 4.3, 4.13, 4.23, 4.25, 4.39, 4.47
Graded: 4.4, 4.14, 4.24, 4.26, 4.34, 4.40, 4.48
```

4.4 Heights of adults.

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.



Min	147.2
Q1	163.8
Median	170.3
Mean	171.1
SD	9.4
Q3	177.8
Max	198.1

Figure 1:

(a) What is the point estimate for the average height of active individuals? What about the median? (See the next page for parts (b)-(e).)

Answer:

The point estimate for the average height of active individuals is 171.1.

The point estimate for the median height of active individuals is 170.3.

(b) What is the point estimate for the standard deviation of the heights of active individuals? What about the IQR?

Answer:

The point estimate for the standard deviation height of active individuals is 9.4.

The point estimate for the IQR height of active individuals is IQR = Q3 - Q1 = $177.8 - 163 = 14$.

(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.

Answer:

In order to be considered unusually tall, that is than an adult has to have greater than 2 standard deviations above the norm; in this particular case, we can take the $Z \geq 2$.

And by solving the equation

$$Z = \frac{x - \mu}{\sigma}$$

Solving for 180 cm.

```
x <- 180
mu <- 171.1
sigma <- 9.4
z <- (x - mu)/sigma
```

Since z = 0.9468085 and this result is less than 2, we conclude that this is considered not unusual.

Solving for 155 cm.

```
x <- 155
mu <- 171.1
sigma <- 9.4
z <- (x - mu)/sigma
```

Since z = -1.712766 and $|z|$ is less than 2, we conclude that this is considered not unusual.

(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.

Answer:

Based on the above, I would not expect the expert to obtain the same values as the ones indicated in the chart. The values that I would expect to see from the new sample would be similar but not necessarily the same.

(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 $SD_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$)? Compute this quantity using the data from the original sample under the condition that the data are a simple random sample.

Answer:

For this case we will use the *Standar Error*.

```
n <- 507
SE <- sigma/sqrt(n)
```

The *Standar Error* for the above will be SE = 0.42.

4.14 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.

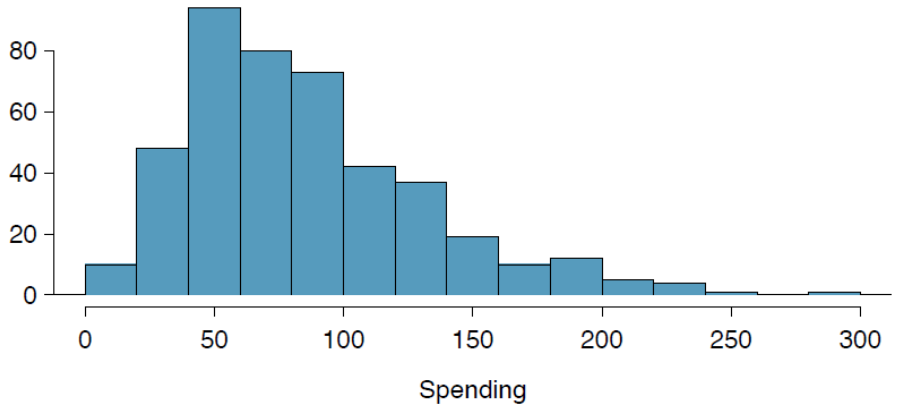


Figure 2:

(a) We are 95% confident that the average spending of these 436 American adults is between \$80.31 and \$89.11.

Answer:

Base on the above, I would conclude that the statement is **False**.

Reason is as follows: From the above explanation, we know 100% that the average spending costs of the sampled American adults is between \$80.31 and \$89.11. We have to remember that the point estimate is always in the confidence interval if it was obtained from a population, in this case our population is 436 American adults.

(b) This confidence interval is not valid since the distribution of spending in the sample is right skewed.

Answer:

Base on the above, I would conclude that the statement is **False**.

Reason is as follows: We already have $n = 436$ and $n \geq 30$ hence the skew does not play an important role for this sample size.

(c) 95% of random samples have a sample mean between \$80.31 and \$89.11.

Answer:

Base on the above, I would conclude that the statement is **False**.

Even thought those intervals would contain the actual mean, we can not assure that exactly 95% of random samples will have a sample mean in 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.

Answer:

Base on the above, I would conclude that the statement is **True**.

By definition of a Confidence Interval this is true.

(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.

Answer:

Base on the above, I would conclude that the statement is **True**.

If we do not need to be as accurate, then we can lower or confidence interval; this will result in a narrowing the confidence interval, making it more inexact.

(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.

Answer:

Base on the above, I would conclude that the statement is **False**.

From our formula we have as follows:

$$SE = \frac{\sigma}{\sqrt{n}}$$

$$SE = \frac{\sigma}{3 \cdot \sqrt{n}}$$

$$SE = \frac{\sigma}{\sqrt{3^2 \cdot n}}$$

$$SE = \frac{\sigma}{\sqrt{9 \cdot n}}$$

If we want to reduce our margin of error three times we will have to increase our sample size 9 times.

(g) The margin of error is 4.4.

Answer:

z = 1.96

n = 436

mu = 84.71

sigma = *unknown*

$SE = \frac{\sigma}{\sqrt{n}}$

LowerTail = mu - z * SE

UpperTail = mu + z * SE

ME <- z * SE

Since Margin Error = $z \cdot SE$ or Margin Error = (Upper Tail - Lower tail)/2 we obtain as follows:

```
UpperTail <- 89.11
LowerTail <- 80.31
ME <- (UpperTail - LowerTail)/2
```

The margin of error is 4.4.

Base on the above, I would conclude that the statement is **True**.

4.24 Gifted children, Part I

Researchers investigating characteristics of gifted children collected 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 distribution of the ages (in months) at which these children first counted to 10 successfully. Also provided are some sample statistics.

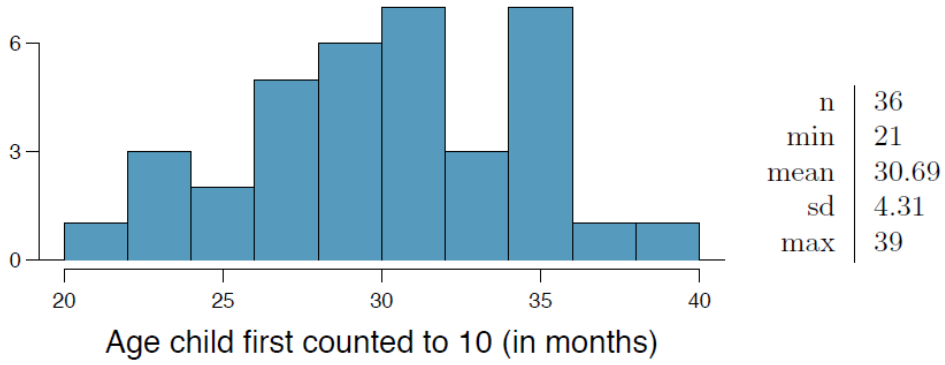


Figure 3:

(a) Are conditions for inference satisfied?

Answer:

Since this data was collected from children in a large city, we can assume that the sample data is likely independent. Also, since n is considered large enough, it satisfies the minimum n needed. Based on the above graph, it seems that there's a normal distribution shape since it doesn't appear to be an obvious skew.

Based on the above, I will conclude that the conditions for inference are 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 first count to 10 successfully is less than the general average of 32 months. Use a significance level of 0.10.

Answer:

```
x <- 32
n <- 36
min <- 21
mu <- 30.69
sigma <- 4.31
max <- 39
SignificanceLevel <- 0.10
```

Null hypothesis (H0): The average development for a child is $\mu = 32$ months.

Alternate hypothesis (HA): The average development for a child is $\mu \neq 32$ months.

Let's consider this as a two side tailed test.

```
SE <- sigma/sqrt(n)
Z <- (mu - x)/(SE)
p <- pnorm(Z, mean = 0, sd = 1) * 2
```

Is the **p-value** = 0.0682026 equal than the significance level? **FALSE**

Hence we **reject** the Null hypothesis (H0).

(c) Interpret the p-value in context of the hypothesis test and the data.

Answer:

Since we ha a **p-value** lower than the significance level, this will suggest that these gifted children count to 10 faster than a normal child.

(d) Calculate a 90% confidence interval for the average age at which gifted children first count to 10 successfully.

Answer:

Since this is a two tailed, I will take the **z** value for 95% to compensate for the lower 5% making it 90%.

```
LowerTail <- round(mu - 1.645 * SE,2)
UpperTail <- round(mu + 1.645 * SE,2)
```

The 90% Confidence interval is from 29.51 to 31.87.

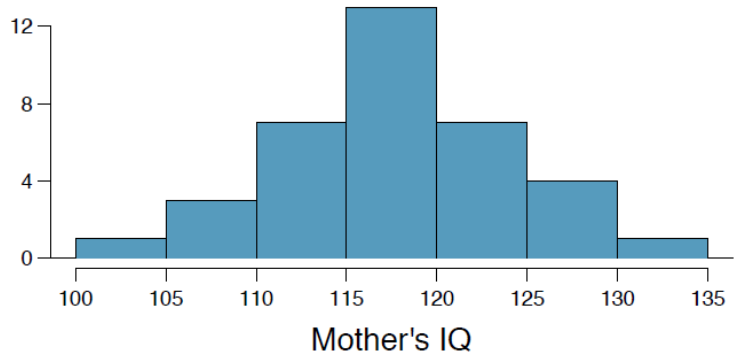
(e) Do your results from the hypothesis test and the confidence interval agree? Explain.

Answer:

Yes, these results agree since we have already **rejected** the Null hypothesis by taking $\mu = 32$.

4.26 Gifted children, Part II

Exercise 4.24 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

Figure 4:

(a) Perform a hypothesis test to evaluate if these data 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.

Answer:

```
x <- 100
n <- 36
min <- 101
mu <- 118.2
sigma <- 6.5
max <- 131
SignificanceLevel <- 0.10
```

Null hypothesis (H0): The average of mother's IQ of gifted children = population's IQ average.

Alternate hypothesis (HA): The average of mother's IQ of gifted children $\neq$ population's IQ average.

Since we don't know the μ and σ values for the population, and by assuming that that the mothers are independent of each other; and since n is big enough, we know that the above case comply with the rules for the inference are satisfied, hence we can utilize the values from the sample as a whole for the population.

Let's consider this to be a two side tailed test.

```
SE <- sigma/sqrt(n)
Z <- (mu - x)/(SE)
p <- (1 - pnorm(Z, mean = 0, sd = 1)) * 2
```

Is the **p-value** = 0 equal than the significance level? **FALSE**

Hence we **reject** the Null hypothesis (H0) and **accept** the Alternate hypothesis (HA).

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

Answer:

Since this is a two tailed, I will take the **z** value for 95% to compensate for the lower 5% making it 90%.

```
LowerTail <- round(mu - 1.645 * SE,2)
UpperTail <- round(mu + 1.645 * SE,2)
```

The 90% Confidence interval is from 116.42 to 119.98.

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

Answer:

Yes, these results agree since we have already **rejected** the Null hypothesis; that is that our **p-value** = 0, which is less than the given significance level of 0.10. and by looking at our confidence interval, these values do not include the given 100 in between, thus agreeing with our alternate hypothesis.

4.34 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.

Answer:

The sampling distribution of the mean is the mean of the population from which the scores were sampled. Therefore, if a population has a mean μ, then the mean of the sampling distribution of the mean is also μ.

4.40 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?

Answer:

```
mu <- 9000
sd <- 1000
x <- 10500
p <- round(1 - pnorm(x, mean = mu, sd = sd),4)
```

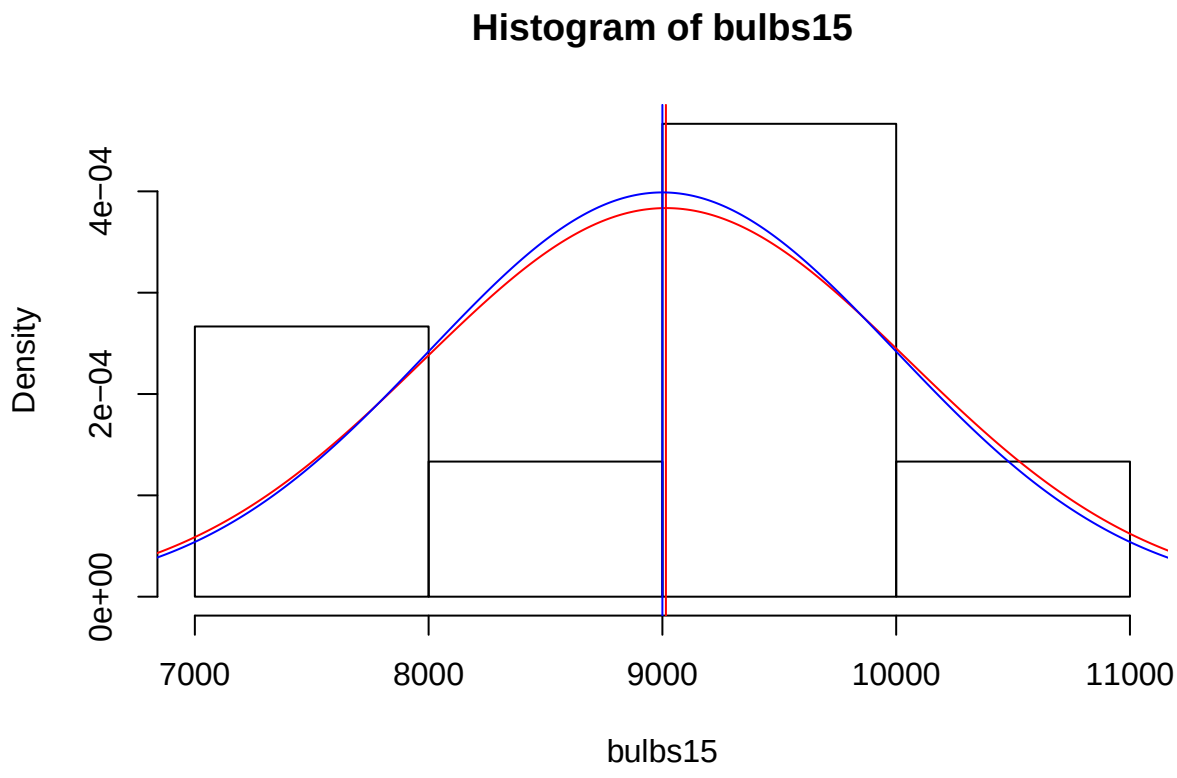
The probability that a randomly chosen light bulb lasts more than 10,500 hours is 6.68%.

(b) Describe the distribution of the mean lifespan of 15 light bulbs.

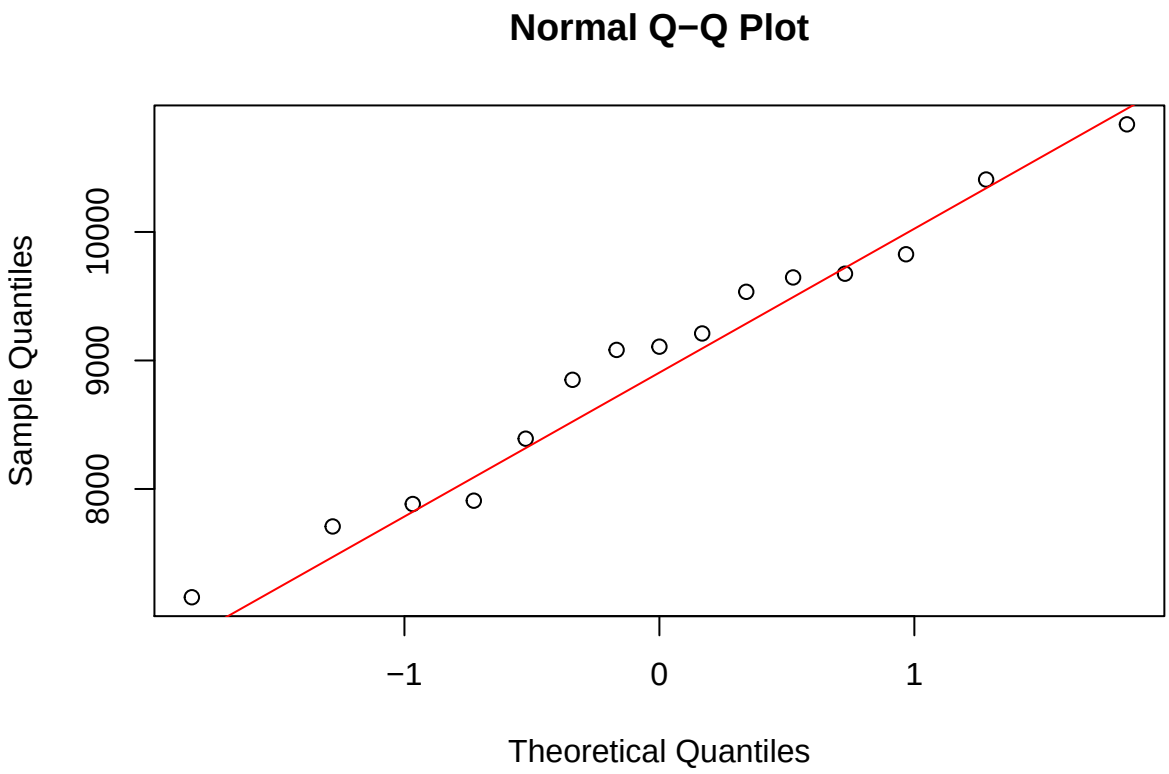
Answer:

I have transposed the normal distribution functions **blue** as the one containing the original mean = 9000 and the **red** one containing the new μ for the $n = 15$.

```
n <- 15
bulbs15 <- rnorm(n, mean = mu, sd = sd)
mu15 <- mean(bulbs15)
sd15 <- sd(bulbs15)


hist(bulbs15, probability = TRUE)
x <- 0:15000
y15 <- dnorm(x = x, mean = mu15, sd = sd15)
y <- dnorm(x = x, mean = mu, sd = sd)
lines(x = x, y = y15, col = "red")
abline(v=mu15,col="red")
lines(x = x, y = y, col = "blue")
abline(v=mu,col="blue")
```




```
qqnorm(bulbs15)
qqline(bulbs15, col = 2)
```



Based on the visual representation, we can make a determination that both distributions are similar but not equal. That is that our $n = 15$ is considered small in order to start approximating our normal distribution in **blue**.

(c) What is the probability that the mean lifespan of 15 randomly chosen light bulbs is more than 10,500 hours?

Answer:

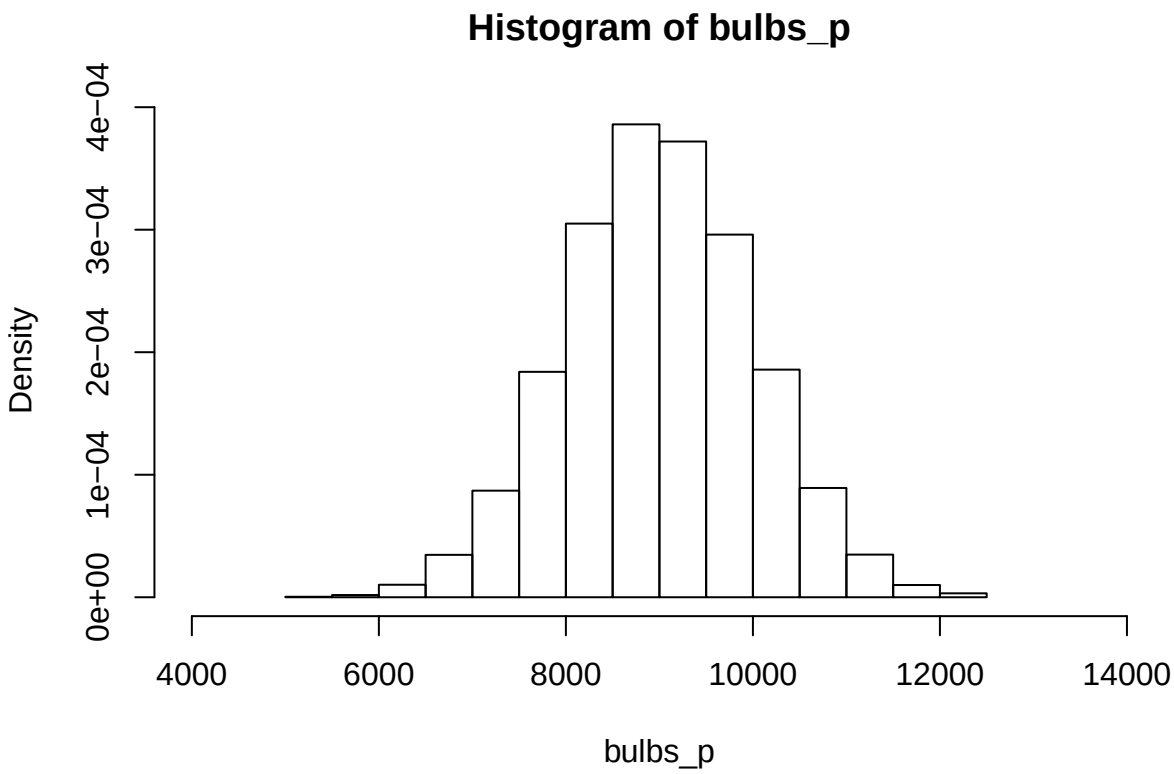
```
n <- 15
x <- 10500
mu <- 9000
sd <- 1000

SE15 <- sd15/sqrt(n)
p15 <- round((1 - pnorm(x, mean = mu, sd = SE15)) * 100,4)
```

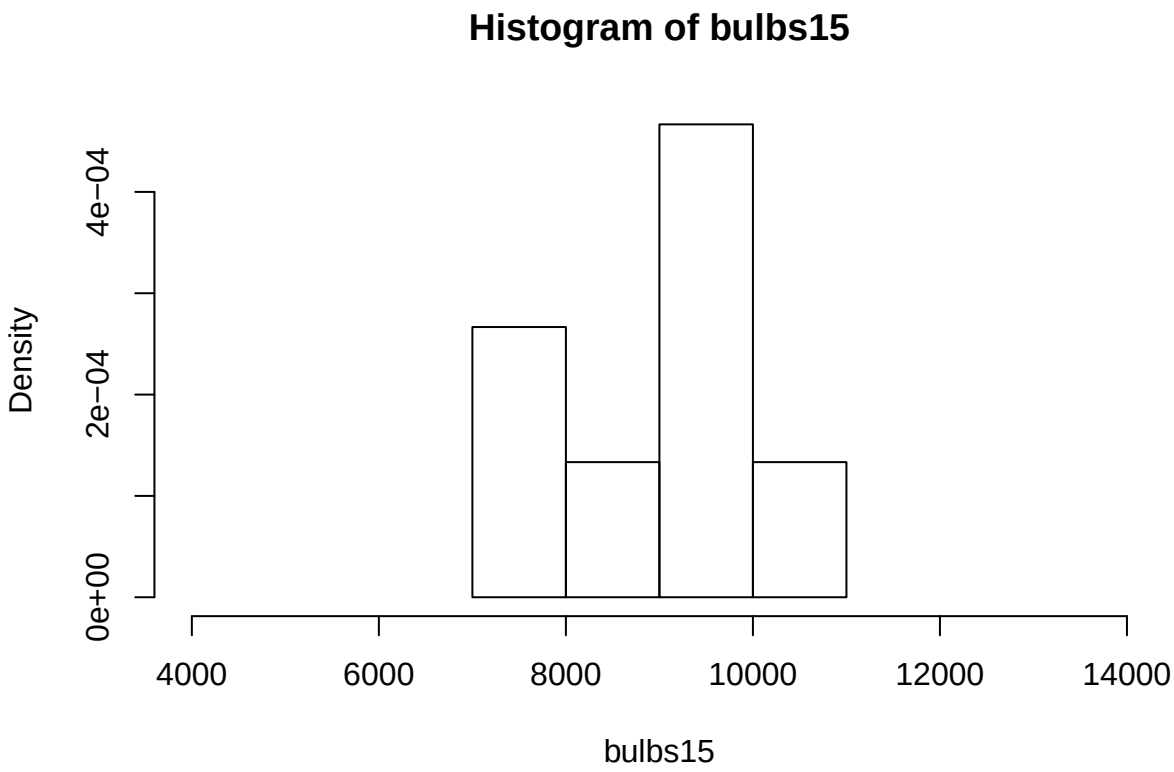
The probability that the mean lifespan of 15 randomly chosen light bulbs is more than 10,500 hours is approximately 0%.

(d) Sketch the two distributions (population and sampling) on the same scale.

Answer:

For this, let's take a big enough population p, let's say p = 10000


Histogram of bulbs_p
Density
0e+00
1e−04
2e−04
3e−04
4e−04
4000
6000
8000
10000
12000
14000
bulbs_p



(e) Could you estimate the probabilities from parts (a) and (c) if the lifespans of light bulbs had a skewed distribution?

Answer:

If you have a skewed distribution, we can not estimate the probabilities as one of the assumptions in order to perform these calculations is that there must be a normal distributions.

4.48 Same observation, di???erent 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.

Answer:

Let's calculate both cases:

n1 = 50

n2 = 500

Since we don't know the σ value for the population, we can still apply as follows: σ = unknown; we can do some work as follows:

$$SE = \frac{\sigma}{sqrtn}$$

If we do the calculation for

$$SE_1 = \frac{\sigma}{sqrtn1} > SE_2 = \frac{\sigma}{sqrtn2}$$

That is $SE_1 > SE_2$

Now the we know the above result, we can continue with the Z value calculation:

$$Z = \frac{\mu - x}{SE}$$

That is

$$Z_1 = \frac{\mu - x}{SE_1} < Z_2 = \frac{\mu - x}{SE_2}$$

And by calculating the respective probabilities, we find that the p value will always change if the sample size changes. in the above case the p-value will decrease based in the final formula of (1 - pnorm(Z, mean = 0, sd = 1)).