CUNY MSDA 606 Spring 2016 Ch. 4 Graded Problems

James Topor

Chapter 4 Problems 4.4, 4.14, 4.24, 4.26, 4.34, 4.40, 4.48

4.4

a) The point estimate is the mean, which is 171.1. The median is 170.3

b) The standard deviation for the point estimate is 9.4. The IQR is 177.8 - 163.8 = 15

c) Is 180cm unusually tall? To find out, we'll calculate a Z score:

```
(180 - 171.1) / 9.4
```

```
## [1] 0.9468085
```

The Z score of .946 is signicantly less than 1.96 (the Z score equivalent to a distance of 2 standard deviations from the mean). As such, a height of 180cm would not be considered to be unusually tall for this population.

Is 155cm unusually short? Again, compute the Z score:

```
(155 - 171.1) / 9.4
```

```
## [1] -1.712766
```

The Z score of -1.71 is also less than 2 standard deviations away from the mean so we conclude that a height of 155cm is not unusually short for this population.

d) We should not expect that a second sample yields a mean and standard deviation identical to the ones given since sampling only provides an approximation of the mean and standard deviation and variation can occur as a result of the sampling process.

e) We use the standard error to calculate the variability of a sample, calculated as sd / sqrt(n). So for our sample distribution we have 9.4/sqrt(507):

```
9.4/sqrt(507)
```

```
## [1] 0.4174687
```

= .4174

4.14

n = 436; mu = 84.71; 95% conf int = (80.31, 89.11)

a) False - the confidence interval applies to the entire population and not these 436 specific individual samples.

b) False- While the distribution appears to be strongly right skewed, the relative large sample size of 436 is sufficient for us to use a normal approximation when there are prominent outliers present.

c) False. The confidence interval applies to the population, not the sample means.

d) True - the confidence interval applies to the population.

e) True - a 90% confidence interval would be narrower since it would not extend as far from the mean as does the 95% confidence interval.

f) False - When calculating a standard error, we divide the standard deviation by the square root of the sample size. To cut the SE (or margin of error) to 1/3 its size, we would need to sample 3^2 = 9 times the number of people in the initial sample.

g) True. Our 95% confidence interval allows us to compute the margin of error as either (89.11 - 80.31) / 2 or (89.11 - 84.71):

```
(89.11 - 80.31) / 2
```

```
## [1] 4.4
```

4.24

n = 36; mu = 30.69; sd = 4.31

a) Are conditions for inference met? Explain any assumptions:

1. Samples must be independent events: TRUE
2. Samples must be randomly selected: TRUE
3. If sampling without replacment, sample size must be less than 10% of the population: TRUE (assuming the city has at least 360 gifted children who've recently turned four years of age)
4. The distribution of original population cannot be heavily skewed. TRUE

b) Hypothesis test w/ .10 signficance level: is the age 32 months or less?

We can state the required hypothesis test as follows: H0: mu = 32 HA: mu < 32

Since we are testing whether the mean is LESS THAN a specific value for our HA, we have a one-sided (left tail) hypothesis. Since we are required to use a .10 significance level, our Z-score cutoff value will be -1.28, i.e., any Z-score less than -1.28 will provide compelling evidence in support of HA.

We can compute the Z-score as follow:

Z = 30.69 - 32 / SE =

```
# compute the standard error
se = 4.31/sqrt(36)

# now find the Z score
(30.69 - 32) / se
```

```
## [1] -1.823666
```

Given a sample mean of 30.69, We have a Z score of -1.82. Since -1.82 is less than the cutoff value for our one-sided hypothesis test, we reject the null hypothesis. Therefore, 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.

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

Now find the p-value associated with a Z score of -1.82:

```
pnorm(-1.82)
```

```
## [1] 0.0343795
```

Since the P-value of 0.034 is less than the significance level of 0.10, we reject the null hypothesis and conclude that we have sufficient evidence to claim that the mean age at which gifted children count to 10 is less than the general average of 32 months.

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

30.69 +- 1.645 * SE =

```
se = 4.31 / sqrt(36)
lower = 30.69 - 1.645 * se
upper = 30.69 + 1.645 * se
c(lower, upper)
```

```
## [1] 29.50834 31.87166
```

The 90% confidence interval is (29.50, 31.87).

e) The results of the confidence interval confirm the results of the hypothesis test since the sample mean (30.69) is within the bounds of the confidence interval. As such, we were correct in our rejection of H0.

4.26

n = 36; mu = 118.2; sd = 6.5

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.

H0: mu = 100
HA: mu != 100 (implies 2 sided test since problem doesn't ask us to prove if mothers of gifted children have an IQ that is GREATER than the average IQ for the population at large).

Using a significance level of 0.10, we have Z-score cutoff values of (-1.64 <= Z <= 1.64), i.e., any Z score outside of this range will indicate that H0 should be rejected.

We can compute our Z score as follows:
Z = 118.2 - 100 / SE =

```r
# compute standard error
se = 6.5 / sqrt(36)

# compute Z score
(118.2 - 100) / se
```

```
## [1] 16.8
```

We have a Z score of 16.8, which is far outside the cutoff range. Therefore, we should reject the null hypothesis H0 and conclude that the IQ of mothers of gifted children is likely different from that of the population at large.

We can also examine the p-value for a Z value being (-16.8 <= Z <= 16.8):

```r
2 * (1 - pnorm(16.8))
```

```
## [1] 0
```

The p-value = ZERO, which is less than the significance value of .10, so we reject H0 and conclude that the IQ of mothers of gifted children is likely different from that of the population at large.

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

118.2 +- 1.645 * SE =

```r
se = 6.5 / sqrt(36)
lower = 118.2 - 1.645 * se
upper = 118.2 + 1.645 * se
c(lower, upper)
```

```
## [1] 116.4179 119.9821
```

90% confidence interval = (116.41, 119.98)

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

Yes the results from the hypothesis test and the confidence interval agree since the value from H0 (mu = 100) lies far outside the confidence interval. As such, we were correct in rejecting H0.

4.34 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.

The sampling distribution represents the distribution of the point estimates based on samples of a fixed size from a certain population. The distribution of a sampling distribution of the mean is approximately normal. The approximation can be poor if the sample size is small, but it improves with larger sample sizes.

Per the Central Limit Theorem we are told that when drawing a single random sample, the larger the sample is the closer the sample mean will be to the population mean. In other words, We increase precision by increasing the number of samples we take. As such, variation is reduced and the standard deviation necessarily decreases in magnitude as we increase the size of the sample.

As the sample size increases, the sampling distribution gets more unimodal, symmetric, and approaches normality. The variability also decreases since the denominator in the standard error calculation is the square root of the sample size. As such, variation is reduced and the standard deviation necessarily decreases in magnitude as we increase the size of the sample.

4.40

mu = 9000; sd = 1000

a) What is the probability that a randomly chosen light bulb lasts more than 10,500 hours?

Find Z score: (10500 - 9000) / 1000

```
zScore <- (10500 - 9000) / 1000
1 - pnorm(zScore)
```

```
## [1] 0.0668072
```

So the probability that a randomly chosen light bulb lasts more than 10,500 hours is **.066**

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

The sampling distribution of the population is known and the data are nearly normal, so the sample mean will also be nearly normal and the distribution will be:

N(mu, SE) where SE = sd/sqrt(n)

```
se <- 1000/sqrt(15)
se
```

```
## [1] 258.1989
```

So we have N(mu = 9000; SE = 258.20).

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

Calculate Z using SE instead of sd:

```
zScore <- (10500 - 9000) / se
zScore
```

```
## [1] 5.809475
```

```
1 - pnorm(zScore)
```

```
## [1] 3.133452e-09
```

So the probability is approximately ZERO.

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

NOTE: This plot makes use of the sample code provided in the slides from our third Meetup/Lecture. The R Markdown file containing that sample code can be found at the following Github link:

https://github.com/jbryer/IS606Spring2016/blob/master/Slides/2016-02-18-Distributions.Rmd

Specifically, lines 177 - 194 of that file were used as the basis for the R code used here to create the required plot.

The author of the R Markdown file is:

title: "IS606 - Distributions" author: Jason Bryer, Ph.D. date: February 18, 2016

The sample code was modified to conform to the parameters of this specific homework problem.

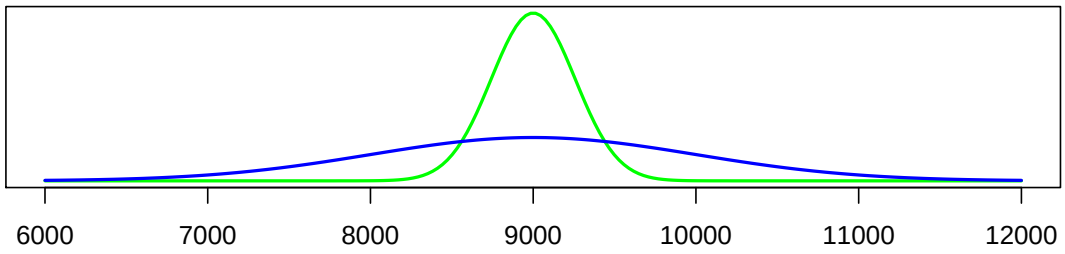
```
# set dimension/boundaries of plotting space
par.orig <- par(mfrow=c(1,2), mar=c(2,1,1.5,1))
par(par.orig)

# set the width of the X axis
x <- seq(6000,12000,length=200)

# set the parameters for the sample relative to the y axis
y <- dnorm(x,mean=9000, sd=se)

# plot the distribution of the sample
plot(x, y, type = "l", lwd = 2, xlim = c(6000, 12000), ylab='', xlab='Sample Distribution = GREEN, Popul

# set the parameters for the population relative to the y axis
y <- dnorm(x,mean=9000, sd=1000)
# plot the distribution of the population
lines(x, y, type = "l", lwd = 2, xlim = c(6000, 12000), ylab='', xlab='', yaxt='n', col='blue')
```



Sample Distribution = GREEN, Population Distribution = BLUE

```
# return to normal 1x1 plotting:
par(mfrow = c(1, 1))
```

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

a) = No because we wouldn't be able to use the normal approximation function to calculate the probability.

b) = No because the sample size of 15 is less than the minimum of 30 needed to avoid skewing problems.

4.48

The P value will decrease. As sample size increases, the standard error will decrease since the denominator in the standard error formula is sqrt(n).

As the standard error decreases, the corresponding Z value will increase, which serves to decrease the p value.

To summarize: As N increases => SE decreases => Z increases => P decreases