4.4

(a)

The point estimate for the average height is equal to the sample mean, which is 171.1 centimeters.

The point estimate for the median height is equal to the sample median, which is 170.3 centimeters.

(b)

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

The IQR = Q3 – Q1 = 14

(c)

For 180 cm, $Z = \frac{180-171.1}{9.4} = 0.947$, which is not unusually tall since it is within 1 SD of the mean.

For 155 cm, $Z = \frac{155-171.1}{9.4} = -1.713$, which is not unusually short since it is within 2 SD of the mean.

(d)

No. The mean and the standard deviation of the new sample will be slightly different from the ones given above. The parameters of each sample are estimates of the population parameters, and they vary from one sample to another because of randomness.

(e)

We use the SE to quantify the variability of such an estimate, which is $\frac{9.4}{\sqrt{507}} = 0.417$ for this sample's mean.

4.6

(a)

We are building a distribution of sample statistics, in this case the sample mean. Such a distribution is called a sampling distribution.

(b)

Because we are dealing with the distribution of sample means, we need to check to see if the Central Limit Theorem applies. The sample size is less than 30, so it does not meet the demand of Central Limit Theorem. The distribution of sample means may not be normal. I think the shape of this distribution could be symmetric, right skewed or left skewed. There is not enough information to judge. It could be symmetric if the 100 times of sampling are randomly and widely selected, and people within a sample are not biased and their attitudes are independent.

(c)

Because we are dealing with a sampling distribution, we measure its variability with the standard error.

SE = $\frac{\sqrt{\frac{14}{15}\times\frac{1}{15}}}{\sqrt{15}} = 0.064$ in approximate calculation. Based on the data of one sample, we assume that the probability p = 14/15, and the standard deviation is equal to $\sqrt{p(1-\text{p})} = \sqrt{\frac{14}{15}\times\frac{1}{15}}$, approximately to Bernoulli distribution.

(d)

The variability of sample means will be smaller with the greater sample size. Because of $SE = \frac{\sigma}{\sqrt{n}}$, if n increases, SE will decrease.

4.8

Recall that the general formula of confidence interval is point estimate $\pm Z \times SE$.

The point estimate is 52%, Z = 2.58 for a 99% confidence level of normal distribution, and SE = 2.4%. So the 99% confidence interval is $52\% \pm 2.58 \times 2.4\%$. That is (45.8%, 58.2%).

We are 99% confident that the fraction of U.S. adult Twitter users who get some news on Twitter is between 45.8% and 58.2%.

4.10

(a)

False. With the significance level of 0.01, the confidence interval is (45.8%, 58.2%). There exists probability that less than half U.S. adult Twitter users get some news on Twitter. So the data does not provide significant evidence that more than half do.

(b)

False. The standard error means the error of the point estimate. In this case, it is the error of estimate for the value 52%. It does not imply how many users were included in the study.

(c)

False. If we want to reduce the SE, we should collect more data to enlarge the sample size. $SE = \frac{\sigma}{\sqrt{n}}$.

To increase the value of n can reduce the value of SE.

(d)

False. For a 90% confidence interval, the range will be wider than a 99% confidence interval. The significance level is greater for a 90% interval. So the Z score will be smaller. The general formula of confidence interval is point estimate $\pm Z \times SE$. So the 90% interval will not be wider. Since we are less confident for an estimate, its corresponding interval should be narrower.

4.12

(a)

We are 95% confident that US residents have stress, depression, problems with emotions, and other not good mental health for 3.40 to 4.24 days during the past 30 days in 2010.

(b)

"95% confident" means that the probability of the average number of not good mental health days during the past 30 days in 2010 for all the US residents between 3.40 and 4.24 is 95%.

(c)

The 99% confidence interval will be larger than the 95% confidence interval. Since we are more confident for an estimate, its corresponding interval should be wider. The significance level is smaller for a 99% interval. So the Z score will be larger. The margin of error will also be larger.

(d)

The SE will be larger. Because $SE = \frac{\sigma}{\sqrt{n}}$, the sample size n decreases, and the SE will increase.

4.14

(a)

False. Inference is made on the population parameter, not the point estimate. The point estimate is always in the confidence interval.

(b)

False. Provided the data distribution is not very strongly skewed, and the sample size of 436 is relatively large enough, the confidence interval is still valid. The Z parameter for a non-normal distribution is greater than that of normal distribution. So the confidence interval should be valid.

(c)

False. The confidence interval is not about a sample mean. It is for an inference of the population.

(d)

True. The confidence interval is for an inference of the whole population.

(e)

True. The 90% confidence interval will have a smaller Z* when computing the margin of error. So the 90% confidence interval will be narrower than that of 95%.

(f)

False. In the calculation of margin of error, the formula is $Z^* \times SE = Z^* \times \frac{\sigma}{\sqrt{n}}$. The Z remains the same for a 95% confidence interval. In order to get a third of the original margin of error, we would need to use a sample $3^2 = 9$ times of the initial one.

(g)

True. The margin of error is half the width of the interval.

4.16

Independence: sample from < 10% of population, and it is a random sample. We can assume that the women in this sample are independent of each other with respect to number of exclusive relationships they have been in. The sample size is much more than 30. The right skew is slight, and the sample is very large so we can definitely make an inference of the population. For a 95% confidence interval, the margin of error is $2\times SE = 2\times\frac{\sigma}{\sqrt{n}} = 2\times\frac{4.72}{\sqrt{5534}} = 0.13$. So the interval is (23.31, 23.57). We are 95% confident that the average age at first marriage of women in US is between 23.31 and 23.57.

4.18

(a)

$H_0 : \mu = 1100$ (The average calorie intake of a diners at this restaurant remains to be 1100 calories.)

$H_A : \mu \neq 1100$ (The average calorie intake of a diners at this restaurant is different from 1100 calories.)

(b)

$H_0 : \mu = 462$ (The average GRE Verbal Reasoning score has not changed since 2004. It remains to be 462.)

$H_A : \mu \neq 462$ (The average GRE Verbal Reasoning score has changed since 2004.)

4.20

The hypotheses should be about the population mean (μ), not the sample mean ($\bar{x}$). The alternative hypothesis implies that it is a one-sided test. However, the social scientist is interested in both increase and decrease of the age. So the test should be two-sided. Correction:

$H_0 : \mu = 23.44$ years old

$H_A : \mu \neq 23.44$ years old

4.22

(a)

Her claim is not well supported. Because $100 is not in the confidence interval. The claim could be wrong.

(b)

A 90% confidence interval will be narrower than the 95% confidence interval. So $100 will be further away from the 90% confidence interval. Her claim will be unreasonable based on the 90% CI.

4.36

(a)

Left skewed. There is a long tail on the lower end of the distribution but a much shorter tail on the higher end.

(b)

I expect most students to have scored above 70 points, as the median is greater than the mean in this left skewed distribution.

(c)

We should not, since it is not a symmetric normal distribution.

(d)

Even though the population distribution is not normal, the conditions for inference are reasonably satisfied, with the possible exception of skew. If the skew isn't very strong (we should ask to see the data), then we can use the Central Limit Theorem to estimate this probability. For now, we will assume the skew isn't very strong. The sample size is also greater than 30. So we can use normal distribution to describe the sample mean. Use $N(70, \frac{10}{\sqrt{40}})$: $P(\mathrm{X} > 75) = 0.00078$

```
> pnorm(75, mean = 70, sd = 10/sqrt(40), lower.tail = FALSE)
[1] 0.0007827011
```

(e)

Standard deviation of the mean is $\frac{\sigma}{\sqrt{n}}$. It would increase to $\sqrt{2}$ of the initial one.

4.40

(a)

```
> pnorm(10500, mean = 9000, sd = 1000, lower.tail = FALSE)
[1] 0.0668072
```

$P(\mathrm{X} > 10500) = 0.067$

(b)

The population SD is known and the data are nearly normal, so the sample mean will be nearly normal with distribution $N(\mu, \frac{\sigma}{\sqrt{n}})$ i.e. $N(9000, \frac{1000}{\sqrt{15}}) = \mathrm{N}(9000, 258.2)$.

(c)

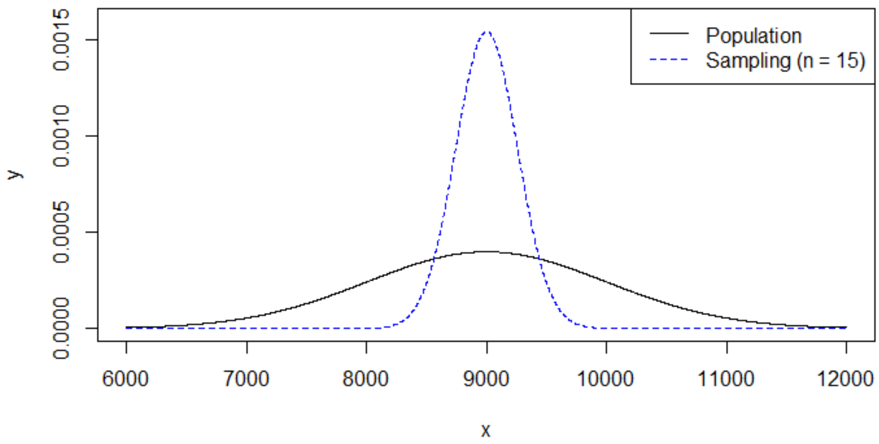
```
> pnorm(10500, mean = 9000, sd = 1000/sqrt(15), lower.tail = FALSE)
[1] 3.133452e-09
```

The probability is nearly 0.

(d)

```
> x = seq(6000, 12000,length = 10000)
```

```
> y = dnorm(x,9000,1000)
> plot(x,y,col="black",ylim=c(0,0.0016),type='l')
> lines(x,dnorm(x,9000,1000/sqrt(15)),col="blue",lty=2)
> legend("topright",legend=paste(c("Population","Sampling (n = 15)")),lwd=
1, lty=c(1,2), col=c("black","blue"),text.font = 1.5)
```



(e)

We could not estimate (a) without a nearly normal population distribution. We also could not estimate (c) since the sample size is not large enough to yield a nearly normal sampling distribution if the population distribution is not nearly normal.

4.44

(a)

$H_0 : \mu = 8\%$ (The 8% of nearsightedness of children is accurate.)

$H_A : \mu \neq 8\%$ (The 8% of nearsightedness of children is inaccurate.)

(b)

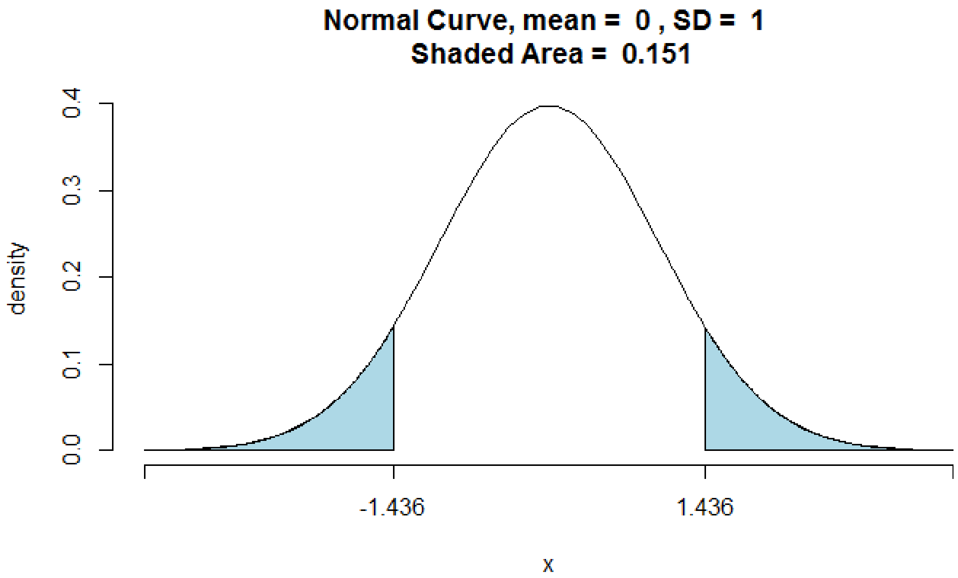
The proportion of nearsighted children in this sample is $21/194 = 0.108$.

(c)

$$Z = \frac{\bar{x} - \mu}{SE} = \frac{0.108 - 0.08}{0.0195} = 1.436$$

(d)

```
> library("tigerstats", lib.loc="C:/Program Files for operation/R-3.3.1/li
brary")
> pnormGC(bound = c(-1.436,1.436), region = "outside", mean = 0, sd = 1, g
raph = TRUE)
[1] 0.1510023
```



The p-value is 0.151 for this hypothesis test.

(e)

Because the p-value is larger than 0.05, we are failed to reject the null hypothesis. The conclusion is that null hypothesis is right. i.e. The 8% of nearsightedness of children is accurate.

4.48

Because of $Z = \frac{\overline{x}-\mu}{SE} = \frac{\overline{x}-\mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\overline{x}-\mu}{\sigma}\sqrt{n}$, if n is larger, $|Z|$ will be larger too. Then the area that constitutes p-value will decrease. So p-value will decrease.