Emma Erdtmann, Daniel Detore
I pledge my honor that I have abided by the Stevens Honor System.

Problem 1

(i)

X is the MPG values from a normally distributed population.

(ii)

```r
# X is the mpg values
X <- c(41.5,50.7,36.6,37.3,34.2,45.0,48.0,43.2,47.7,42.2,43.2,44.6,48.4,46.4
,46.8,39.2,37.3,43.5,44.3,43.3,35.8,33.9,40.1,41.3,37.7,39.6,42.4,41.7,35.7)

X_sd <- 3.15
X_bar <- mean(X)
X_n <- length(X)
X_alpha <- 0.05

X_Z <- 1 - X_alpha/2
X_margin <- qnorm(X_Z)*(X_sd / sqrt(X_n))
X_ci_1 <- c(X_bar - X_margin, X_bar + X_margin)
print((paste("Confidence interval known sd @ 95% manually calculated:",
X_ci_1)))

# I could only figure out the built in formula for CI of unknown population
standard deviation
```

```
[1] "Confidence interval known sd @ 95% manually calculated:
40.632848393817"
[2] "Confidence interval known sd @ 95% manually calculated:
42.9257722958382"
```

(iii)

```r
X_t <- qt(X_Z, df = X_n-1)*(X_sd/sqrt(X_n))
X_ci_2 <- c(X_bar - X_t, X_bar + X_t)
print((paste("Confidence interval unknown sd @ 95% manually calculated:",
X_ci_2)))

X <- list(mpg = X)
model <- lm(mpg ~ 1, X)
X_ci_2 <- confint(model, level=0.95)
print((paste("Confidence interval unknown sd @ 95% using R function:",
X_ci_2)))

```

```
[1] "Confidence interval unknown sd @ 95% manually calculated:
40.5811144255297"
[2] "Confidence interval unknown sd @ 95% manually calculated:
42.9775062641254"
[1] "Confidence interval unknown sd @ 95% using R function:
40.0651167550972"
[2] "Confidence interval unknown sd @ 95% using R function:
43.493503934558"
```

(iv)

```r
print(paste("Margin of error for 1ii (z):", X_Z))
print(paste("Margin of error for 1iii (t):", X_t))
```

```
[1] "Margin of error for 1ii (z): 0.975"
[1] "Margin of error for 1iii (t): 1.19819591929786"
```

The confidence interval where the population standard deviation is known would be the most accurate, which is confirmed by the fact that the confidence interval & margin of error in 1ii is smaller than in 1iii.

(v)

```r
u_bound <- qchisq(X_alpha/2, df = X_n - 1)
l_bound <- qchisq(1 - X_alpha/2, df = X_n - 1)
X_ci_3 <- c((X_n - 1)*X_sd^2 / l_bound, (X_n - 1)*X_sd^2 / u_bound)
print((paste("Confidence interval of the population variance @ 95% manually calculated:", X_ci_3)))
```

```
[1] "Confidence interval of the population variance @ 95% manually
calculated: 6.24887656123692"
[2] "Confidence interval of the population variance @ 95% manually
calculated: 18.1494990136156"
```

Problem 2

(i)

Y is if MPG > 40, where 1 = yes, 0 = no.

(ii)

```r
Y <- ifelse(X > 40, 1, 0)
sample <- Y[1:16]
p <- mean(sample)
n <- length(sample)
margin <- qnorm(1 - (1 - 0.925)/2) * sqrt((p * (1-p)) / n)
Y_ci <- c(p - margin, p + margin)
print(paste("Confidence interval @ 92.5%:", Y_ci))
```

```
[1] "Confidence interval @ 92.5%: 0.557259081195296"
[2] "Confidence interval @ 92.5%: 0.942740918804704"
```

$$E = z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 1.780464\sqrt{\frac{0.75(1-0.75)}{16}} = 0.1927409$$

$$\hat{p} \pm E = 0.75 \pm 0.1927409 = (0.5572591, 0.9427409)$$

(iii)

$$E = z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 1.780464\sqrt{\frac{0.6551724(1-0.6551724)}{26}} = 0.1571495$$

$$\hat{p} \pm E = 0.6551724 \pm 0.1571495 = (0.4980229, 0.8123219)$$

```
> p2 <- mean(mpg)
> E <- qnorm(1 - (1 - 0.925)/2) * sqrt((p2 * (1 - p2)) / length(mpg))
> print(c(p2 - E, p2 + E))
[1] 0.4980229 0.8123219
```

(iv)

$E_{ii} = 0.1927409$

$E_{iii} = 0.1571495$

The process in iii, where we use all of the data points, is more accurate because its error margin is smaller. This means, if we want to draw any conclusions about the data set, we should use all of the data available.

Problem 3

(i)

```
> library(BSDA)
> x = rnorm(30, mean = 1, sd = 2)
> z.test(x, mu = 1, sigma.x = 1, conf.level = .95)

      One-sample z-Test

data:  x
z = 1.5752, p-value = 0.1152
alternative hypothesis: true mean is not equal to 1
95 percent confidence interval:
 0.929761 1.645439
sample estimates:
mean of x
   1.2876
```

(ii)

```
> N <- 0
> for (x in 1:100) {
+     x <- rnorm(30, mean = 1, sd = 2)
+     low  <- mean(x) - qnorm(1 - .05 / 2) * 2 / sqrt(30)
+     high <- mean(x) + qnorm(1 - .05 / 2) * 2 / sqrt(30)
+     if (low < 1 && 1 < high) {
```

```
+            N <- N+1
+      }
+ }
> print(N/100)
[1] 0.96
```
The proportion is approximately equal to the confidence level.

(iii)
```
> x = rnorm(30, mean = 1, sd = 2)
> t.test(x, mu = 1, conf.level = .90)

       One Sample t-test

data:  x
t = 0.70471, df = 29, p-value = 0.4866
alternative hypothesis: true mean is not equal to 1
90 percent confidence interval:
 0.5941651 1.9810346
sample estimates:
mean of x
   1.2876
```

(iv)
```
> N <- 0
> for (x in 1:100) {
+      x <- rnorm(30, mean = 1, sd = 2)
+      low  <- mean(x) - qt(1 - .10 / 2, 29) * 2 / sqrt(30)
+      high <- mean(x) + qt(1 - .10 / 2, 29) * 2 / sqrt(30)
+      if (low < 1 && 1 < high) {
+            N <- N+1
+      }
+ }
> print(N/100)
[1] 0.89
```
The proportion is approximately equal to the confidence level. This is possible because the Student's T distribution is an approximation of the normal distribution, which the data actually is. It's made more accurate by our use of the correct amount of degrees of freedom (here $df = n - 1$).