Problem 1: What value of $z^*_{\alpha/2}$ or $t^*_{\alpha/2}$ is used to construct:

a. A 95% confidence interval to estimate μ if the sample size is 23. (Assume random sample assumption and normal population distribution assumption are met.)

```
qt(1-0.05/2,df=23-1)
```

Answer: $t^*_{\alpha/2} = 2.074$

b. A 98% confidence interval to estimate μ if the sample size is 1982. (Assume random sample assumption is met.)

```
qt(1-0.02/2,df=1982-1)
```

Answer: $t^*_{\alpha/2} = 2.328$

c. A 98% confidence interval to estimate p if the number of successes is 991 and the number of failures is 991 in a random sample.

```
qnorm(1 - 0.02/2)
 # OR
qnorm(0.99)
```

Answer: $z^*_{\alpha/2} = 2.326$

d. Why are the quantiles for parts b and c close in value? Give a reason other than that they both have the same degree of confidence, 98%, and the sample sizes are the same.

Answer: Because for very large n (specifically if n is larger than 30), $t^*_{\alpha/2} \approx z^*_{\alpha/2}$

Problem 2: Getting2NoU data - Mean Exercise (hr/-day)

This time, we will use the Getting2NoU data to construct a confidence interval to estimate students' average daily amount of time spent exercising.

a. Construct a histogram and Q-Q plot of daily exercise, exercise. Include both plots in your submission. Describe the shape of the distribution.

```
> summary(exercise)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   1.000   1.500   1.729   2.000  11.000
```

Answer: The shape of the distribution is right-skewed as displayed by both the histogram and Q-Q plot.

b. What is the point estimate of population mean exercise, μ?

```
> mean(exercise)
[1] 1.728557
```

Answer: $\hat{\mu} = 1.729$ hours/day

c. What is the standard error of the sample mean for exercise?

```
> sd(exercise)
[1] 1.511372
> length(exercise)
[1] 335
```

Answer: $SE = \frac{s}{\sqrt{n}} = \frac{1.511372}{\sqrt{335}} = 0.0826$

d. Write R code to construct a 95% confidence interval for R using the mean, standard error and sample size from previous parts of this problem.

```
> # By hand 95% CI construction
> # c(-1,1) serves as -/+ in the formula
> mean(exercise) + c(-1,1)*qt(0.975,length(exercise)-1)*
        sd(exercise)/sqrt(length(exercise))
[1] 1.566124 1.890990
```

e. Use t.test() to compute the 95% confidence interval to estimate the population mean daily exercise. Interpret the result.

```
> t.test(exercise, conf.level=0.95, alternative="two.sided")
```

```
One Sample t-test
data: exercise
t = 20.933, df = 334, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 1.566124 1.890990
sample estimates:
mean of x
 1.728557
```

Interpretation: We are 95% confident that the mean number of hours a student spends exercising is between 1.57 to 1.89 hours/day.

f. What are the assumptions / conditions we rely on to construct a confidence interval for mean μ? Do you think the data meets these assumptions? Do you think the result is reliable? Explain.

Answer: When we use the t-distribution to construct a confidence interval for μ, we assume that the sample is randomly drawn from some population that is approximately normal.

The t distribution for computing the sample size is robust to non-normality when $n \geq 30$ and no outliers are present. The distribution of the sample is right-skewed. The sample size is 335 which is much greater than 30, therefore by the CLT we can assume that the sample mean is distributed normally in the absence of outliers. A boxplot of the data indicates that five outliers are present. The results based on the t distributions may not be reliable in this case due to the presence of five outliers.

Problem 3: State H_0 and H_a

For each problem, decide on the population parameter, direction and value being tested.

a. A pharmaceutical company claims that its new flu vaccine is effective for at least 85% of people.

Answer:
$H_0 : p = 0.85$
$H_a : p > 0.85$

b. A city planner believes that the average commute time in the city is more than 30 minutes.

Answer:
$H_0 : \mu = 30$
$H_a : \mu > 30$

c. A company claims that 90% of its customers are satisfied with their service.

Answer:
$H_0 : p = 0.90$
$H_a : p \neq 0.90$

d. A teacher wants to know if the average score on a final exam is different from 75.

Answer:
$H_0 : \mu = 75$
$H_a : \mu \neq 75$

e. A factory claims that less than 5% of its products are defective.

Answer:
$H_0 : p = 0.05$
$H_a : p < 0.05$

HW6 Problem 4: Planning the Sample Size for a Confidence Interval on μ

A biologist is studying the average cholesterol level (in mg/dL) in the blood of a certain animal species. From a small pilot study, the sample standard deviation of cholesterol was about 18 mg/dL. For a larger upcoming study, the biologist wants to plan the sample size needed to estimate the true mean cholesterol level μ with 95% confidence and with a margin of error no greater than 4 mg/dL.

a. State the formula used for planning the sample size. Which quantities depend on the confidence level?

Answer: When the population standard deviation is unknown but estimated by s^*, a common planning formula is:

$$n = \left(\frac{z^* s^*}{m}\right)^2$$

Here, z^* depends on the desired confidence level (for 95%, $z^* = 1.96$), while s^* is estimated from prior data and m is the target margin of error.

b. Compute the minimum required sample size using $s^* = 18$, $m = 4$, and $z^* = 1.96$.

Answer:

$$n = \left(\frac{1.96 \times 18}{4}\right)^2 = 77.8$$

Always round up to the next integer, so the minimum sample size is: $n = 78$.

c. Interpret the result in context, using our target confidence level and target moe.

Answer: The biologist must collect data from at least 78 animals to ensure that a 95% confidence interval for the mean cholesterol level has a margin of error no larger than 4 mg/dL.